The similarity approach strikes back: Negation in counterfactuals[1]

Katrin SCHULZ — *University of Amsterdam*

Abstract. The meaning of counterfactual conditionals is standardly described using the similarity approach (Stalnaker, 1968; Lewis, 1973). This approach has recently been challenged by Ciardelli et al. (2018). They argue that the similarity approach is in principle unable to account for the meaning of counterfactuals with an antecedent consisting of a conjunction embedded under a negation ($\neg(p \wedge q)$). Ciardelli et al. (2018) dismiss the approach on these grounds and offer an alternative. The main goal of the present paper is to defend the similarity approach against this attack. I will argue that the problem that underlies the observations in Ciardelli et al. 2018 is more general and not solved by the solution they offer. I will furthermore argue, against Ciardelli et al. (2018), that the cause of the problem is not the similarity approach, but the interaction of negation with the meaning of counterfactual conditionals. The paper will conclude with a first outline of a solution for the problem, which still uses the similarity approach, but combines it with an alternative semantics for negation.

Keywords: counterfactuals, negation, similarity approach, causality.

1. Introducing the main players and the storyline

How should we approach the semantics of counterfactual conditionals? If you look at the literature on this topic over the last 50 years, you will see that there is one particular approach that clearly dominates the field: the similarity approach of Stalnaker (1968) and Lewis (1973). We teach it to our students the first time they encounter the problem of counterfactual sentences and they grow up under the impression that this is the only way one should think about them. It became a paradigm, an empire in the vast field of the literature on counterfactuals. But paradigms come with a serious drawback: they can make us blind. We start to mistake theory for reality and, consequently, don't question it anymore. That also seemed to happen in the case of the similarity approach. Even though at the beginning the approach was challenged from various angles, the criticism dried out as the approach became more and more established.

However, in a recent paper by Ciardelli et al. (2018) the similarity approach was called into question again. A team of Skywalkers stepped forward and challenged the empire. They put forward an argument that targets the very core of the approach and claim that this argument convincingly shows that we need to give up our paradigm, dismiss the similarity approach. In this paper we will take the side of the empire and pick up the glove that has been thrown at its feet. We will argue that even though the argument of Ciardelli et al. (2018) is extremely valuable, it does not succeed in eliminating the similarity approach. There is a way to account for the observations they make without giving up the paradigm.

We will start in Section 2 with a short introduction to the similarity approach and premise semantics for counterfactuals. In Section 3 we will have a look at the recent challenge brought

[1]I would like to thank Ivano Ciardelli, Luca Champollion and the audience at Sinn und Bedeutung 22 for feedback and discussion. Special thanks to Jonathan Pesetsky for proof-reading the manuscript.

© 2018 Katrin Schulz. In: Uli Sauerland and Stephanie Solt (eds.),
Proceedings of Sinn und Bedeutung 22, vol. 2, ZASPiL 61, pp. 343–360. ZAS, Berlin.

forward by Ciardelli et al. (2018). We will discuss their evidence against the similarity approach and the alternative approach they propose. In Section 4 we will present our evidence against their proposal. We will argue that this evidence points actually to a more general problem concerning the interpretation of negation in conditionals. An alternative solution for the problem is sketched in Section 5. Section 6, contains conclusions and an outlook on future work.

2. The galactic empire

2.1. The similarity approach

From the perspective of possible worlds, the central question any approach to the meaning of counterfactual conditionals has to answer is the question of the selection function. A counterfactual is true if in a selected set of possible worlds that make the antecedent true, the consequent is true as well.[2] But which situations should be selected? As Goodman (1955) has shown, it cannot be the set of all possible worlds that make the antecedent true. The conditional (1) seems intuitively to be true. But the consequent of the counterfactual doesn't hold in all possibilities that make the antecedent true. What, for instance, if the match had been soaked in water overnight? This example shows that when we evaluate a counterfactual, we consider only a particular subset of the antecedent worlds. But how to select the right worlds?

(1) If I scratched this match, it would light.

The core idea of the similarity approach is that we select the possible worlds in which the antecedent is true and which in other respects differ minimally from the evaluation world w_0 of the counterfactual. This idea can be made precise using an order over possible worlds that, given the actual world, compares all other worlds with respect to their similarity to the actual world. This order is at least assumed to be a weak total order that centers around the actual world w_0 (the actual world is a smallest element of the order). A counterfactual with antecedent A and consequent C is now said to be true in case the consequent holds in all possible worlds that make the antecedent true and are minimal with respect to the order.[3]

There exist various refinements of this theory, imposing all kinds of extra conditions on the order. The argument against the similarity approach that will be discussed in the next section targets the basic core of the theory, which is what we outlined here.

2.2. Premise semantics

We can also take an inferential perspective on the truth conditions of counterfactuals. Then we could say that a counterfactual is true in case we can infer the consequent from the antecedent. From the inferential perspective, the question of the selection function discussed above becomes the questions of the premise function. It is not possible to infer the consequent just from

[2]This set can consist of one or more worlds, depending on the theory.

[3]For the purpose of this paper we follow Stalnaker (1968) and adopt the Limit Assumption.

the antecedent. Certain facts of the evaluation world are used as additional premisses of this inference. To infer the consequent of (1) from its antecedent, we need to take into account the laws governing the behaviour of matches. We also assume (because this is true for the match in front of me) that the match wasn't soaked in water overnight. In premise semantics this is spelled out in terms of the *premise set P*. *P* is the set of true facts of the evaluation world that matter for the truth of a counterfactual. A counterfactual is said to be true in case the consequent can be inferred from the antecedent together with the laws and any maximal subset of the premise set consistent with the antecedent. Choosing maximal subsets consistent with the antecedent makes sure that we take as many premises into account as possible, without running into a contradiction. Let Π be a set of sentences. We define $Max_{\Pi}(\phi)$ as the set of maximal subsets of Π consistent with ϕ. Then we can define the truth conditions of a counterfactual $A \rightsquigarrow C$ according to premise semantics as in A (Veltman, 1976; Kratzer, 1981b, a).

$$A \rightsquigarrow C \text{ iff } \forall S \in Max_P(A) : S \cup \{A\} \models C. \tag{A}$$

Suppose, for instance, the premise set P consists of the sentences p, q, and r and we want to evaluate a counterfactual with the antecedent $\neg p$. The unique maximal subset of P consistent with the antecedent would be the set $\{q, r\}$. A counterfactual with the antecedent $\neg p$ is true, in case the consequent follows from $\neg p$ together with q and r (and the relevant laws). It might happen that there are multiple equally maximal subsets of the premises that are consistent in the antecedent. In this case Clause A demands that the consequent has to follow from each of them together with the antecedent. Consider, for instance, a counterfactual with the antecedent $\neg p \vee \neg q$ using the same premise set. In this case there are two equally maximal subsets of P that are consistent with the antecedent: $\{p, r\}$ and $\{q, r\}$. Rule A now demands that both of these sets together with the laws and the antecedent entail the consequent.

2.3. The relation between similarity approach and premise semantics

If you think about it, premise semantics is actually not that different from the similarity approach discussed before. The premises that together with the antecedent have to entail the consequent characterise the relevant antecedent worlds that we need to check for the truth of the consequent. Also in case of premise semantics, we want these selected worlds to be as close as possible to the actual world; we want to keep as many of the premises as possible. We can define an order on possible worlds that compares them with respect to the premises they make true: given the premises P we say that a world w_1 is more similar to the actual world w_0 than a world w_2 in case the subset of P true in w_2 is a subset of the subset of P true in w_1. Based on this order the similarity approach will make the same predictions as Rule A.[4] Going back to our example with the premise set $\{p, q, r\}$, this set would induce the order on possible worlds given in the left diagram of Figure 1 (for each world only those premises are given that are true in this world, false premises are left out). The worlds w_3, w_5, w_6 and w_7 all make the antecedent

[4]For the formal details see Lewis 1981. If restrict ourselves to similarity relations that are strict partial orders, the equivalency also holds the other way around: given a similarity order, one can define a premise set P such that Rule A counts the same counterfactuals true. We can, thus easily switch from one perspective to the other.

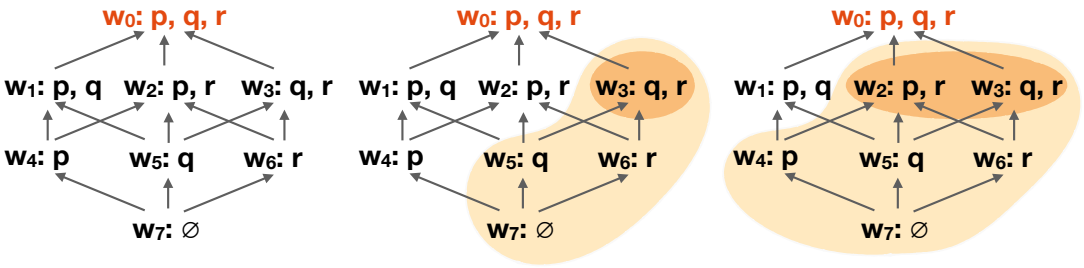

Figure 1: Evaluating counterfactuals given the premise set $\{p, q, r\}$.

If $\neg p$ *had been the case* true. Among these, w_3 is the world most similar to the actual world w_0 (dark orange in the second diagram of Figure 1). This is also the world where the maximal subset of the premises consistent with the antecedent is true. A conditional with antecedent $\neg p \vee \neg q$ is true in the worlds w_2, w_3, w_4, w_5, w_6 and w_4. The worlds most similar to the actual world are w_2 and w_3 (dark orange in the right diagram of Figure 1). They correspond to the two maximal subsets of the premises consistent with the antecedent that we calculated before.

This finishes our short presentation of the current paradigm for how to approach the meaning of conditional sentences. This is the empire in our story. Both perspectives, the similarity approach and premise semantics, will play a role in the discussion below. The attack of Ciardelli et al. (2018) is directed against the formulation using a similarity order, but for their alternative approach Ciardelli et al. (2018) build on premise semantics.

3. The empire under attack

3.1. Earlier strikes at the empire

We mentioned already at the beginning that the similarity approach has been attacked before. However, it is quite hard to really falsify the proposal. The reason is its generality. The argument has to work for any possible similarity order. It has to hit the very idea of approaching the meaning of counterfactuals using an order relation on possible worlds.

One way to truly hit the approach is by targeting its logic. The semantics of the similarity approach can be axiomatized (Lewis, 1973). The axioms capture the meaning of counterfactuals in terms of the inferences you are allowed to draw with them. One could attack the approach by arguing that the axioms the similarity approach give rise to are not the right ones: important properties of counterfactuals are not covered or some of the predicted inferences are in fact not valid for counterfactuals. An example for such an attack is the discussion concerning the law *Simplification of Disjunctive Antecedents* (SDA), see formula B. This law is not valid according to the logic of the similarity approach. In other words, SDA is not entailed by the axiomatisation. However, the principle seems to be intuitively valid, not only for counterfactuals (2a), but for conditionals in general (2b). Therefore, it has been argued, B should be a law of any adequate theory of the meaning of counterfactuals. The similarity approach doesn't tick this box, hence, the argument continues, we need a different approach.

$$(SDA) \quad [(\phi \vee \psi) \rightsquigarrow \chi] \rightarrow [(\phi \rightsquigarrow \chi) \wedge (\psi \rightsquigarrow \chi)] \tag{B}$$

(2) a. If Mary or Sue had been at the party, it would have been a lot more fun.
b. If it's sunny tomorrow or aliens invade Amsterdam overnight, I will eat breakfast outside.
c. If Spain had fought with the Axis or the Allies, she would have fought with the Axis.

This line of attack is not without problems. Some authors have argued that, while (SDA) holds for the normal resolution of similarity, it is not generally valid. See, for instance, examples as in (2c): from this counterfactual one cannot infer that if Spain had fought with the Allies, it would have fought with the Axis. But this wouldn't get the similarity approach completely off the hook; one would still need an account of the normal resolution of similarity. A different way to counter this attack is by replying that it only shows that the logic of the similarity approach needs to be strengthened. In other words, we need to put extra conditions on the similarity relation. However, there is an extra complication here. One can prove that no compositional account of the meaning of counterfactuals based on classical logic can validate (SDA) without validating *Antecedent Strengthening* (AS), given in formula C.

$$(AS) \quad [\phi \rightsquigarrow \chi] \rightarrow [(\phi \wedge \psi) \rightsquigarrow \chi] \tag{C}$$

Now, we certainly don't want (AS) to hold for the meaning of counterfactuals. This was the point of example (1): from *If I scratched this match, it would light* it doesn't follow *If the match was soaked in water overnight and I scratched it, it would light*. On the one hand, this sounds like bad news for the similarity approach. It clearly shows that we cannot account for (SDA) by strengthening the logic.[5] But you could also take this to be good news. The result shows that the validity of (SDA) is not a particular problem of the similarity approach. It is a problem of any approach to the meaning of counterfactuals that involves classical logic. This weakens the power of (SDA) as an argument against the similarity approach in particular. But if we want to adopt the similarity approach, we still need to explain why (SDA) seems intuitively valid.

So far we have been focusing exclusively on the conditional connective as an operator occurring in B. We implicitly assumed that it is the logic of this operator that needs to account for the critical observation. But there is another operator present in the relevant counterfactual: disjunction. Maybe the semantics assumed for the conditional connective is not the problem, but the semantics we assumed for disjunction. There are various other contexts in which the classical approach to disjunction is known to be problematic (Free Choice phenomena, exhaustive interpretation). This is also the angle from which Ciardelli et al. (2018) approach the problem of (SDA).[6] To deal with the semantics of disjunction properly, they propose that we need to work with a more fine-grained semantic framework: inquisitive semantics (Ciardelli et al., 2018). Most importantly, in this framework, the meaning of a sentence is not equated with the

[5] At least not without giving up basic logical principles, like the substitution of logical equivalencies.
[6] They are not the first to do so, see in particular Alonso-Ovalle 2009; Fine 2012; Schulz 2011 for related proposals.

set of worlds in which the sentence is true, but with a set of such sets, representing the maximal information states that would support the sentences. In most cases this set of sets just contains the set of worlds that make the sentence true. But the support condition for disjunctions introduce non-trivial alternatives: for each disjunct the set of worlds that make this disjunct true.[7] The counterfactual operator $\rightsquigarrow$ is then proposed to quantify over the alternatives the antecedent gives raise to, see D below. For the definition of the connective $\mapsto$ you can then pick your favourite notion of counterfactual entailment. It could be a similarity approach, the proposal of Ciardelli et al. (2018), or something else. Whatever you choose, the inference (SDA) will now be valid for $\rightsquigarrow$.

$$\phi \rightsquigarrow \psi \Leftrightarrow \forall p \in Alt(\phi) \exists q \in Alt(\psi) : p \mapsto q \tag{D}$$

Thus, at least in the case of (SDA), what started out as a challenge for the similarity approach eventually led to the development of a more advanced semantics of other operators involved in the critical observation. The similarity approach itself remained relatively unaffected.

3.2. The recent challenge by Ciardelli et al. (2018)

We will now turn to the challenge posed by Ciardelli et al. (2018) for the similarity approach. They also target the logic of the similarity approach. But the critical inference that they address is not one that is invalid according to the similarity approach, but should be valid according to our intuition. In the case of Ciardelli et al. 2018 we are dealing with an inference that is valid according to the logic, but is intuitively invalid according to Ciardelli et al. (2018): the inference in E.

$$[(\neg\phi \rightsquigarrow \chi) \wedge (\neg\psi \rightsquigarrow \chi)] \rightarrow [\neg(\phi \wedge \psi) \rightsquigarrow \chi] \tag{E}$$

Ciardelli et al. (2018) empirically tested the intuitive validity of the inference. They conducted an online experiment in which they asked participants to judge the truth or falsity of the counterfactuals given in (3) in the scenario depicted in Figure 2. In this scenario a circuit connects two switches to a lamp. The wiring is such that the light is on if and only if the switches are in the same position. In the depicted scenario both switches, A and B, are up and the lamp is on.

(3) a. If switch A was down, the light would be off.
b. If switch B was down, the light would be off.
c. If switch A or switch B was down, the light would be off.
d. If switch A and switch B were not both up, the light would be off.

The results of their study are given in Table 1. The important observation is that even though the majority of the participants judged the conditionals (3a) and (3b) to be true, only 22% took

[7] Ciardelli et al. (2018) propose as the support condition of a disjunction $s \models \phi \vee \psi$ iff $s \models \phi$ or $s \models \psi$.

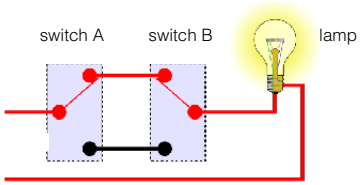

Figure 2: The circuit example of Ciardelli et al. (2018).

sentences	number	true	%	false	%	indet.	%
$\neg A \rightsquigarrow Off$	256	169	66,02%	6	2,34%	81	31,64%
$\neg B \rightsquigarrow Off$	235	153	65,11%	7	2,98%	75	31,91%
$(\neg A \vee \neg B) \rightsquigarrow Off$	362	251	69,33%	14	3,87%	97	26,80%
$\neg(A \wedge B) \rightsquigarrow Off$	372	82	22,04%	136	36,56%	154	41,40%

Table 1: Results of the empirical study.

(3d) to be true as well. However, according to E, if (3a) and (3b) are considered to be true, then (3d) should be true as well. This is a serious problem for the similarity approach. The inference in E is valid for the logic of the similarity approach. That means it holds no matter what similarity relation you choose. Ciardelli et al. (2018) conclude from this that the approach is doomed to fail. The empire falls.

Let us take a closer look at what the problem seems to be. Using the terminology of premise semantics, if (3a) is true, this tells us that the fact that switch B is up is part of the premises of the evaluation world. For the counterfactual to be true, the position of the second switch needs to be kept constant. In the same way the truth of (3b) allows us to conclude that the fact that switch A is up is part of the premises. There might be also other facts that count as premises. We will just consider one other fact, q.[8] The premise set $\{A, B, q\}$ results in the order over possible worlds described in Figure 3, first diagram. The sentence $\neg(A \wedge B)$ is true in the worlds w_1, w_3, w_4, w_5, w_6 and w_7, the area shaded bright orange in Figure 3, second diagram. According to the similarity approach, the most similar worlds are w_1 and w_2 (dark orange in Figure 3, second diagram). In both of these worlds the light is off. Hence, the counterfactual in (3d) is predicted to be true – contra to the results of the empirical study.

The problem seems to be that interpreters of (3d) also consider a world like w_5 where both switches are down. In this world the light is on and, hence, the counterfactual is judged to be false. So, the set that should be selected as the set of relevant antecedent worlds should be the set $\{w_1, w_3, w_5\}$, see the dark orange area in the last diagram of Figure 3. Thus, also worlds not optimal according to the order need to be selected as relevant antecedent worlds.

[8]The reader might wonder why we do not consider the possibility that the state of the lamp is part of the premises. According to Ciardelli et al. (2018) (and many other authors) the reason is that this is a fact causally dependent on the antecedent. Such facts are deselected as possible premisses. But this issue and the way Ciardelli et al. (2018) account for it is completely orthogonal to the topic of this paper. We simply assume that this is taken care of.

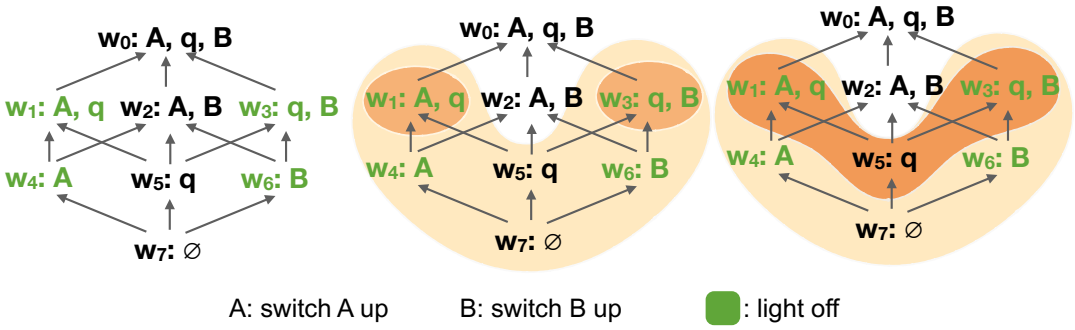

Figure 3: The worlds selected by the similarity approach for the antecedent of Example (3d) (center) and the worlds that should be selected for this antecedent (right).

3.3. The alternative approach of Ciardelli et al. 2018: Cautious retraction

Based on the criticism discussed in the last subsection, Ciardelli et al. (2018) dismiss the similarity approach. They conclude that we need to select the relevant antecedent worlds in a different way. The alternative they propose is spelled out in terms of premise semantics. Recall the interpretation Rule A of standard premise semantics, repeated here as F. According to this rule a counterfactual is true in case all maximal subsets of the premises consistent with the antecedent together with the antecedent entail the consequent.

$$A \mapsto C \text{ iff } \forall S \in Max_P(A) : S \cup \{A\} \models C. \tag{F}$$

Ciardelli et al. (2018) propose to replace this rule with Rule G. According to this rule, a counterfactual is true in case the intersection of all maximal subsets of the premises that are consistent with the antecedent together with the antecedent entail the consequent. They choose to err on the side of caution and only allow fact to be kept constant in case they are part of all maximal subsets consistent with the antecedent.[9] Thus, they predict a smaller subset of the premises to be carried over to the hypothetical scenario considered by the counterfactual, and, as a consequence, less counterfactuals to be true.

$$A \mapsto C \text{ iff } \bigcap Max_P(A) : S \cup \{A\} \models C. \tag{G}$$

With this interpretation rule they can account for observations concerning the critical example (3d). If we assume that the position of the switches, A and B, are part of the premises (together with other facts q), then there are two maximal subsets of the premises consistent with the antecedent $\neg(A \wedge B)$: the sets $\{A,q\}$ and $\{B,q\}$. The intersection only contains q; the positions of both switches in the actual world needs to be given up, because together they contradict the antecedent. We get the correct prediction that the consequent has to be true not only in the worlds w_1 and w_3, but also in w_5. The light isn't off in all of these worlds (not in w_5). Hence, the counterfactual comes out as false, as intended. For the counterfactuals (3a) and (3b)

[9] In fact, they propose that this set sets an upper limit for the premises kept. We will come back to this later.

Ciardelli et al. (2018) make exactly the same predictions as the similarity approach. In these cases there is only one maximal subset of premises that is consistent with the antecedent.

Because the interpretation rule given in G only takes into account the truth-conditions of the antecedent, it predicts identical truth-conditions for counterfactuals with logically equivalent antecedents. Therefore, one might think that this proposal makes wrong predictions for (3c), which has an antecedent that is logically equivalent to the antecedent of (3d). The counterfactual (3c) we do want to come out as true. However, Rule G is combined with Rule D, assuming inquisitive semantics for the treatment of disjunction. From the perspective of inquisitive semantics, while the antecedents of (3c) and (3d) are truth-conditional equivalent, they are not semantically equivalent. Because Rule D is sensible to this semantic difference, we get different truth conditions for the counterfactuals. The counterfactual in (3c) is still predicted to be true. The rule D checks whether each disjunct of the antecedent counterfactually entails the consequent. Whether we define counterfactual entailment using Rule F or Rule G, we obtain that the truth of the consequent is in w_1 and w_3. In these worlds the consequent is true. Hence, the counterfactual is predicted to be true.

4. The empire strikes back–part 1

The empirical results of Ciardelli et al. (2018) seem to be rather devastating for the similarity approach. No matter how the similarity order is defined, there is no way the approach will predict that (3a) and (3b) are true, while (3d) is false. Does this mean that we have to dismiss the approach; give up on the empire? In this section I will argue that this conclusion would be too hasty. First, I will make a more conceptual point and show that the proposal of Ciardelli et al. (2018) can still be seen as an order-based approach. The solution Ciardelli et al. (2018) propose is more a variation of than an alternative to the similarity approach. Secondly, I will claim that the empirical results of Ciardelli et al. (2018) hint at a more general semantic problem. While Ciardelli et al. (2018) are able to account for one particular realisation of this problem, they fail to account for other instantiations. Thus, their solution strategy – targeting the similarity approach – does not seem to work.

4.1. Cautious retraction as cautious similarity

As the authors admit, their proposal comes in spirit very close to premise semantics. But still it is not a standard premise semantics approach. Conceptually, Ciardelli et al. (2018) consider their approach different in that they do not incorporate what they call the minimal change requirement. The central idea is not to keep as many facts of the premises as the antecedent allows, but "... rather, whenever we are faced with a counterfactual assumption, we determine a background of facts which are not at stake, and we hold all these facts fixed." (Ciardelli et al. 2018: 35). The only restriction on the background is that it has to be a subset of $\bigcap Max_P(A)$.[10]

This sounds as if they completely do away with the idea of optimisation in the meaning of counterfactuals. The context fixes some set of background facts, to those facts we add the

[10]See also footonote 8.

antecedent and then we check whether the consequent follows. But that is not a very faithful picture of what is going on here. Looking at their formal apparatus we see two differences from standard premise semantics: (i) the background, the facts relevant for the meaning of a counterfactual, can be a subset of the filtered premisses, and (ii) the condition for how to filter or retract premises has changed. To the first difference a defender of standard premise semantics could reply that premise semantics captures this by making the premise function context dependent.[11] Ciardelli et al. (2018) do not show that their way to incorporate context dependence gives better results. The second difference is more substantial. Ciardelli et al. (2018) could have proposed that the upper limit of the background is the filtered premise set of standard premise semantics: $Max_P(A)$. They opt for being more cautious and choose $\bigcap Max_P(A)$ instead. However, the resulting truth conditions for counterfactuals can still be understood as result of an order-based optimisation process. In other words, and contra to what they seem to say, optimisation still plays a role in the semantics of counterfactuals. We will argue for this by showing that, just as for standard premise semantics, the truth conditions they predict for counterfactuals can be produced by selecting optional worlds based on a similarity order. You only have to be a bit more generous in what you count as an optimal world.

Assume, again, that P is our finite set of premises, the facts of the evaluation word that matter for the truth of a counterfactual.[12] As before, we use P to define a strict partial order on possible worlds: $w_1 \leq_P w_2$ iff $\{\varphi \in P \mid w_2 \models \varphi\} \subseteq \{\varphi \in P \mid w_1 \models \varphi\}$. Let $M_P^+(\varphi)$ be the $<_P$-maxima in the set of worlds that satisfy $\bigcap Max_P(\varphi) \cup \{\varphi\}$. Because P is finite, this set is non-empty. We use $M_P^+(\varphi)$ to define truth conditions for counterfactuals as in A.

$$A \mapsto C \text{ iff } \forall w' : \ [w' \models A \wedge \exists w \in M_P^+(A)(w' \leq_P w)] \rightarrow w' \models C. \tag{H}$$

It can now be shown that the conditions in G and H are equivalent. Thus, to check the truth of a counterfactual, we don't just look at the most similar antecedent worlds, but at all worlds smaller or equal to a certain limit, described by $M_P^+(A)$. Ciardelli et al. (2018) don't give up on similarity, they just relax a bit the order-based selection criterium.

Proof. The result follows from $\{w \mid w \models A \wedge \exists w \in M_P^+(A)(w' \leq_P w)\} = \{w \mid w \models \bigcap Max_P(A) \cup \{A\}\}$. So, we prove this equation.
$\Rightarrow$ Assume $u \in \{w \mid w \models A \wedge \exists w \in M_P^+(A)(w' \leq_P w)\}$. Thus, there exists a world $w \in M_P^+(A)$ such that $v \leq_P w$. Because of the definition of $M_P^+(A)$, it follows that $w \models \bigcap Max_P(\varphi) \cup \{\varphi\}$. Because $v \leq_P w$, it follows $v \models \bigcap Max_P(\varphi)$. We also know that $v \models A$. Thus, $v \in \{w \mid w \models \bigcap Max_P(A) \cup \{A\}\}$.
$\Leftarrow$ Assume $v \in \{w \mid w \models \bigcap Max_P(A) \cup \{A\}\}$. From this it follows $v \models A$. Because P is finite, it follows that there is a maximal w with $v \leq_P w$ and $w \models \bigcap Max_P(A) \cup \{A\}$. Hence, $v \in \{w \mid w \models A \wedge \exists w \in M_P^+(A)(w' \leq_P w)\}$.

[11] Ciardelli et al. (2018) choose for a framework where the premises function is fixed as the set of facts (Ciardelli et al. 2018: 25). Then, context dependence has to be build in at a different place and they choose they notion of background as the right place.

[12] We work with a finite set of premises, because this is also what Ciardelli et al. (2018) do. Additionally, they work with premise sets that consist only of atomic sentences. We don't adopt this restriction here.

4.2. Cautious similarity under scrutiny

We now turn to potential limitations of the alternative proposal of Ciardelli et al. (2018). As noticed before, the interpretation rule that according to Ciardelli et al. (2018) should take over the place of the similarity approach only takes the truth conditions of its arguments into account.[13] Consequently, the approach makes the same predictions for logically equivalent antecedents. It is not that clear that this prediction is actually correct. Take, for instance the antecedent $\neg A$: *"Switch A is down"*. $\neg A$ is logically equivalent to stating that switch A is down and that it is not the case that both switches are up, $\neg A \wedge \neg(A \wedge B)$. We can now compare the truth values assigned to counterfactuals with these two logically equivalent antecedents, see (4a) and (4b). In the scenario in Figure 2 the first counterfactual is dominantly judged to be true (see Table 1). But what about (4b)? Is this counterfactual also intuitively true in the described context? That seems at least questionable. Hence, there appears to be a difference in interpretation of (4a) and (4b). The redundant information $\neg(A \wedge B)$ cannot just be ignored, contra to what the similarity approach and also cautious similarity tell us.

(4) a. If switch A was down, the light would be off.
b. If switch A and switch B were not both up and switch A were down, the light would be off.

One could counter that this is not a particular strong argument against the proposal. Assume that we were to empirically test (4a) and (4b) and observed a significant difference between the truth-judgements of both counterfactuals. It would still be hard to say what caused the difference. Maybe the observed difference is due to pragmatic reasons: the sentence (4b) is reinterpreted because of the redundancies in the antecedent. In other words, we could get rid of the problematic example by moving it to the pragmatic waste basket.

4.3. The limits of cautious similarity–an empirical study

Let's try to make the argument stronger. We also saw that the proposal of Ciardelli et al. (2018) doesn't deviate a lot from the similarity approach. Again, it operates using a set of selected facts of the actual world (premises) that need to be kept true in the selected antecedent worlds. The proposal also tries to keep as many of the premises as possible. The only difference is that Ciardelli et al. (2018) are a bit more cautions about when to keep a premise: only in case this premise is an element of each maximal subset of the premises consistent with the antecedent. So, basically, this is still an approach based on minimisation of differences from the actual world. But if the minimisation forces you to make a choice between two premises, the approach refuses to choose and gives up both.[14] If no such choice needs to be made, the approach makes exactly the same predictions as the similarity approach/premise semantics.

Assume now, I add to my counterfactual antecedent a formula expressing information about the

[13] In case this is not clear already, this holds also for the similarity approach.

[14] Just to compare, standard premise semantics/similarity approach demands that you check the consequent for both choices.

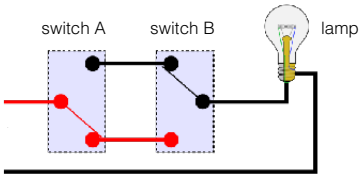

Figure 4: A new scenario with one switch down and no electricity.

premises that is true in the evaluation world. As long as the premises occurring in this formula do not contradict other parts of the antecedent, they will be in each maximal consistent set. Hence, the approach will try to keep them true in the considered counterfactual scenario. Consider, for instance, the counterfactuals in (5) in the scenario described in Figure 4. The wiring is as in the previous scenario of Figure 2, but now the position of the switches is different and we add as an additional fact of the evaluation world that there is no electricity at the moment.

(5) a. If the electricity was working, then then light would be on.
b. If the electricity was working and switch A was up, then the light would be on.
c. If the electricity was working and switch A and switch B were not both up, then the light would (still) be off.

As before, we use A and B to shorten *switch A is up* and *switch B is up* respectively. Additionally, we use E to shorten the claim that the electricity is working. As before, we can use examples like (5a) and (5b) to establish that E, $\neg A$ and B should be among the premises for the scenario presented in Figure 4. Given this premise set, what would Ciardelli et al. (2018) predict for the truth conditions of (5c)? In the first diagram of Figure 5 the different possibilities with respect to this premise set are described (w_0 is, again, the actual world). In the green worlds the light is off. The antecedent of (5c) is true in the worlds w_3, w_5 and w_7. The unique maximal subset of the premises consistent with the antecedent is $\{\neg A, B\}$. Because there is only one maximal subset with the antecedent, the approach makes the same predictions as the similarity approach: w_3 is selected as the world where the consequent needs to be true, marked dark orange in the first diagram of Figure 5. In w_3 the light is off. Thus, the approach predicts the counterfactual to be true.

However, this is not the interpretation that we observe. I conducted an experiment using an online questionnaire, designed with Qualtrics and distributed using Prolifix. The study duplicated the setting of the studies conducted in Ciardelli et al. 2018, only changing the example. Participants were asked to judge the truth/falsity of the counterfactuals given in (5) using a slider bar (see Figure 6). The slider bar allowed for five positions that were in the evaluation translated into the numbers $0-4$. The questionnaire was filled in by 51 native speakers of English, who received 1 Pound as payment. The results are given in Table 2.[15] The first two examples were interpreted in agreement with the predictions of Ciardelli et al. (2018) (and the similarity approach). This also confirms the premise set used to calculate the predictions. However, the

[15]Some of the responses can be questioned, because the participant either answered the fillers incorrectly or finished the study within a few seconds. In row 4 of Table 2 the corrected results are given. They are nearly identical to the unfiltered results.

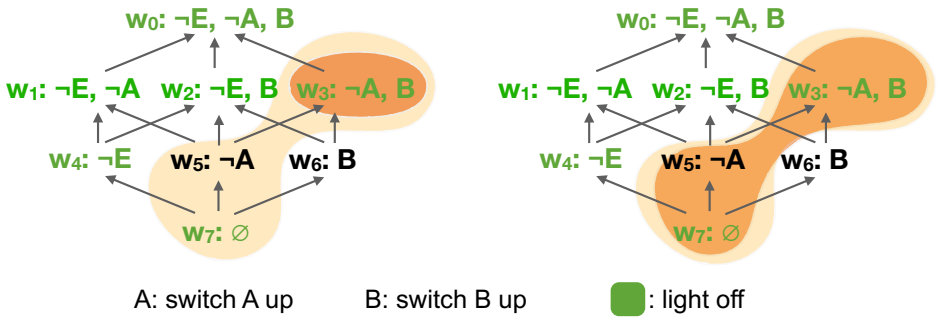

Figure 5: The worlds selected by Ciardelli et al. 2018 for the antecedent of (5c) (left) and the worlds that should be selected for this antecedent (right).

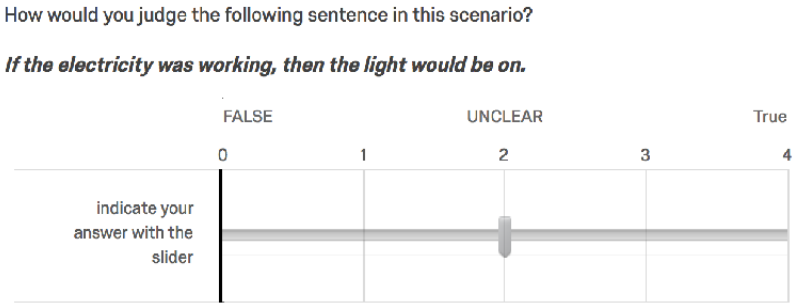

Figure 6: A screenshot from the online questionnaire.

counterfactual (5c) is by the majority of the participants judged to be false. The reason seems to be that the fact that switch B is up and switch A is down in the evaluation world shouldn't be kept constant in the hypothetical scenario introduced by the antecedent. These facts need to be given up. This would give you as the set of antecedent worlds $\{w_3, w_5, w_7\}$, marked dark orange in the second diagram of Figure 5. Then the counterfactual (5c) would correctly be predicted to be false.[16]

There are a couple of options for how Ciardelli et al. (2018) could defend their approach against the results presented here. One way would be to use the same move made in the last subsection and submit the observations to the pragmatic waste basket. In the end, also in this case the antecedent of the critical counterfactual (5c) contains information that is redundant. The only difference is that now the redundancy is context dependent, while before contextual information didn't matter. This is an easy move to be made, but only convincing in case one can back it up with a solid pragmatic story. One would probably want to turn to Grice's maxims of conversation: the speaker must have a reason to include redundant information; this reason is to signal to the speaker that *A*, and *B* shouldn't end up in the premise set, etc. However, making this argumentation explicit won't be easy. There are many possible reasons the speaker might have to mention redundant information. One would still need to explain how the hearer knows

[16]The problem that we noticed here is a problem that approaches to the meaning of counterfactuals based on the similarity approach share. Any approach that allows for substitution of logical equivalents will have trouble accounting for the observations made here.

sentences	true	%	false	%	indet.	%
$E \rightsquigarrow On$	8	16%	42	82%	1	2%
$(E \wedge A) \rightsquigarrow On$	43	84%	5	10%	2	4%
$[E \wedge \neg(A \wedge B)] \rightsquigarrow On$	14	27%	27	53%	8	16%
$[E \wedge \neg(A \wedge B)] \rightsquigarrow On^*$	9	26%	20	59%	5	15%

Table 2: Results of the empirical study.

that excluding A and B from the premise set is the only viable option here. Furthermore, one would have to explain how Gricean reasoning can interact with the operation of a semantic operator (the conditional connective).

4.4. Intermediate conclusions

In this section we explored the limits of the proposal of Ciardelli et al. (2018). We discussed at least one concrete example which the approach cannot immediately account for. We also outlined a possible pragmatic escape route for the approach, but observed that this route needs to be worked out. However, evaluating the proposal of Ciardelli et al. (2018) is not our goal. The purpose of this paper is to defend the similarity approach against the attack of Ciardelli et al. (2018). To some extend we did that in Subsection 4.1 when I argued that the alternative Ciardelli et al. (2018) propose is still an order-based approach and not really giving up on similarity. But also the results of the study conducted can be used to that purpose. They point to a different possible explanation of the data of Ciardelli et al. (2018), in particular one that leaves the similarity approach unaffected.

The antecedent of (5c) is very similar to that of the critical example (3d). Both antecedents involve a complex negation $\neg(A \wedge B)$. In both cases we observe that if we apply minimisation, we lose too many possibilities. In both cases we want to keep – in a certain sense– all logical possibilities that the negation allows. In the next section we want to explore an alternative explanation of the observations made in this paper; one that takes the negation to be responsible instead of the semantics proposed for the conditional. Though, we will not argue here that this solution should be preferred to the proposal of Ciardelli et al. (2018), the fact that this is a plausible alternative explanation of the data shows that we do not need to give up the similarity approach and the empire is safe for now.

5. The empire strikes back–part 2

5.1. ... by blaming negation

In this section I will develop an alternative explanation for the critical data of Ciardelli et al. (2018), one that at the same time can explain the observations made in Section 4. The structure of this solution employs the same strategy that we saw in Section 3.1 in reaction to the observation that the law (SDA) (simplification of disjunctive antecedents) seems intuitively valid for counterfactuals. There, we ended up blaming the disjunction in the antecedent for the validity of the inference. Following Alonso-Ovalle (2009); Fine (2012); Schulz (2011); Ciardelli

et al. (2018) we proposed that the disjunction introduces alternatives for each of its disjuncts. The conditional is then said to quantify over these alternatives, see rule D, repeated here as I. The connective $\mapsto$ that the rule builds on can still be interpreted according to the similarity approach. From a more general point, we analysed the intuitive validity of (SDA) as evidence that we need a richer semantic framework than just basic truth conditions, in particular with respect to the semantic treatment of disjunction.

$$s \models \phi \rightsquigarrow \psi \Leftrightarrow \forall p \in Alt(\phi) \exists q \in Alt(\psi) : p \mapsto q \tag{I}$$

The same solution will be now proposed with respect to the observations of Ciardelli et al. (2018) and Section 4. Again, we take the examples to show that we need a richer semantic framework. In particular, we need to respect the alternatives that expressions might introduce. But in addition to the earlier proposal that disjunction introduces alternatives, we will argue here that this also applies to negation.

5.2. A counterproposal

At the core of the present proposal lies the idea that negation, just as disjunction, introduces alternatives. We already need the semantics of the connective $\rightsquigarrow$ to quantify over alternatives in order to account for disjunctive antecedents. The alternatives that negation gives raise to will be treated the same way. We will argue that this is sufficient to account for the critical observations.

We adopt the framework of inquisitive semantics that Ciardelli et al. (2018) work with.[17] The only thing we need to change is the support condition for negation. The solution we propose is inspired by standard approaches to truthmakers of negations. A truth maker of a formula $\neg\phi$ is standardly taken to be a formula χ that contradicts the formula ϕ in question ($\chi \perp \phi$). We additionally restrict truth makers of negations to relevant sentences/propositions that contradict ϕ. This means we need a notion of relevance here, a question that we want to see answered. As we are concerned with semantics here, we use a notion of relevance that is context independent and relies on the sentence itself. Assuming a propositional language we define $\mathscr{L}(\phi)$ as the set of atomic formula occurring in ϕ. To be relevant according to ϕ is to know the truth value of all elements in $\mathscr{L}(\phi)$. In other words, the question capturing what is relevant according to a sentence ϕ is $Q(\phi)$, the partition introduced by $\mathscr{L}(\phi)$ (i.e. the set of sets of possible worlds that assign the same truth value to all elements in $\mathscr{L}(\phi)$). For example, if $\phi = A \wedge B$, then $\mathscr{L}(\phi) = \{A, B\}$ and $Q(\phi) = \{AB, A\bar{B}, \bar{A}B, \bar{A}\bar{B}\}$.[18] Any formula using the same vocabulary gives rise to the same issue. We extend support to issues in the standard way: an information state s supports an issue I ($s \models I$) in case s completely answers I, i.e. $\exists i \in I : s \subseteq i$. The new interpretation rule for negation is given in J. It states that a situation supports $\neg\phi$ in case it's a complete answer to the issue raised by ϕ and contradicts ϕ.

[17] We could as well have used truthmakers semantics.

[18] To simplify notation we write $A\bar{B}$ to refer to the set of worlds where A is true and B is false.

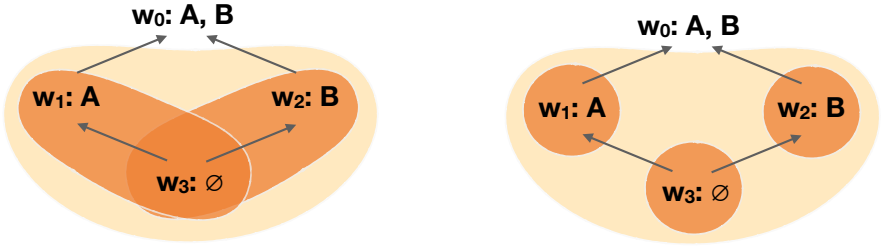

Figure 7: The alternatives predicted for the antecedents of (6a) (left) and of (6b) (right).

$$s \models \neg\phi \text{ iff } s \models Q(\phi) \text{ and } s\bot\phi. \tag{J}$$

According to this rule, the semantic value of the sentences $\neg A \vee \neg B$ and $\neg(A \wedge B)$ differ: $Alt(\neg A \vee \neg B) = \{\bar{A}, \bar{B}\}$ (see the left diagram in Figure 7), but $Alt(\neg(A \wedge B)) = \{\bar{A}B, A\bar{B}, \bar{A}\bar{B}\}$ (see. Crucially, the sentence $\neg(A \wedge B)$ contains an additional alternative, $\bar{A}\bar{B}$. When this sentence occurs as antecedent of a counterfactual, also this alternative needs to counterfactually entail the consequent.

Let us see how this accounts for our examples. First we take a look at the critical examples of Ciardelli et al. (2018). As discussed before, we assume the premises in this case to include the positions of the switches. This gives the order of worlds displayed in Figure 7. The antecedent of (3c), repeated here as (6a) is true in w_1, w_2 and w_3, marked bright orange in the left diagram of Figure 7. The antecedent is disjunctive: $\neg A \vee \neg B$, hence, the counterfactual is predicted to be true if each disjunct separately counterfactually entails the consequent. We employ the similarity approach to compute counterfactual entailment. So, we predict that the counterfactual is true if the consequent is true in world w_1 and w_2 (left diagram of Figure 8). In these two worlds the light is off. Hence, (3c) is correctly predicted to be true.

(6) a. If switch A or switch B was down, the light would be off.
b. If switch A and switch B were not both up, the light would be off.
c. If the electricity was working and switch A and switch B were not both up, then the light would (still) be off.

The negation in the antecedent of (3d), repeated here as (6b), introduces the alternative set given in the right diagram of Figure 7. For each of these alternatives we have to check whether they counterfactually entail the consequent. In this case, this is not true. The alternative set $\{w_3\}$ does not counterfactually entail that the light is off. Hence, the approach correctly predicts that the counterfactual in (6b) is false. Finally the example (6c) in the scenario described in Figure 4. In this case the order over possible worlds looks a bit different, because the facts change, see Figure 9. The alternatives the antecedent gives rise to are $\{w_3\}$, $\{w_5\}$ and $\{w_7\}$. If we now check for each of these alternatives whether it counterfactually entails the consequent, we see that this is not the case. There is one alternative, $\{w_5\}$, that makes the consequent false.

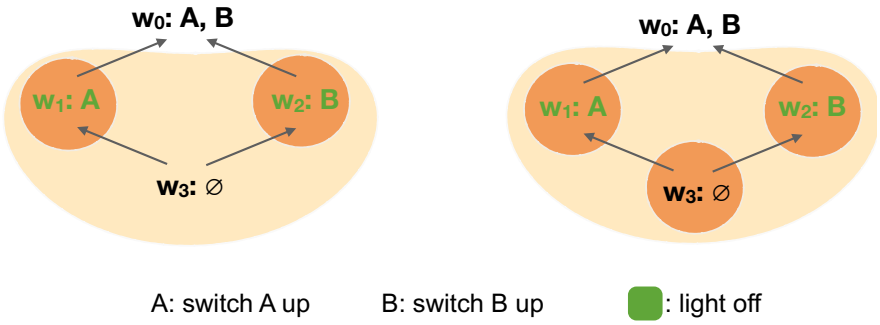

Figure 8: The most similar worlds selected for the antecedents of (6a) (left) and of (6b) (right) assuming rule I and the similarity approach to counterfactual entailment.

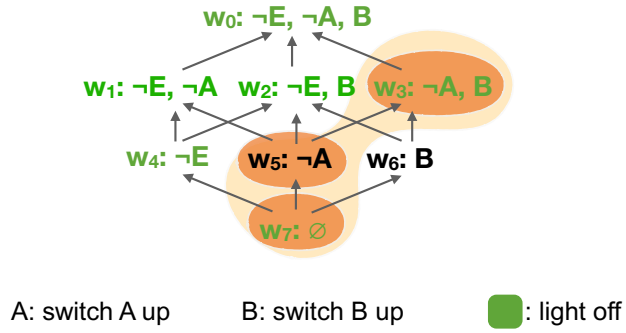

Figure 9: Alternatives predicted for (6c) in the scenario of Figure 4.

Hence, the counterfactual (6c) is predicted to be false, just as intended.

6. Conclusions: The empire is still alive and kicking!

This paper addressed a recent challenge put forward by Ciardelli et al. (2018) against the similarity approach of Stalnaker (1968) and Lewis (1973), the standard approach towards the meaning of counterfactual conditionals nowadays. We have argued that the evidence that Ciardelli et al. (2018) put forward against the similarity approach is not conclusive. Our argument proceeded in two steps. First, we have shown that in certain scenarios also the counter-proposal of Ciardelli et al. (2018) runs into trouble. While their approach can possibly be saved using a pragmatic story, we have sketched an alternative analysis that provides a unified solution for these and the original examples of Ciardelli et al. (2018). This alternative is still compatible with the similarity approach. Hence, the similarity approach is not defeated, yet. The empire is safe.

The solution proposed here builds on inquisitive semantics. We proposed that not only disjunction, but also negation introduces alternatives. The conditional quantifies over these alternatives and checks for each of them separately whether they counterfactually entail the consequent of the counterfactual. We are, then, free to choose our favourite approach to defining this notion of entailment. Nothing stops us from choosing a similarity approach here. As we discussed in the last section, at least for all examples discussed in this manuscript a similarity approach

makes adequate predictions.

Proposing that negation introduces non-trivial alternatives is a big step to take. This step needs to be supported by more evidence, preferably coming from the same sources that motivate the inquisitive treatment of disjunction. The good news is that there is a lot of literature on disjunction that we can build on. But this is work that still needs to be done. Some preliminary independent evidence for the semantics for negation proposed here comes from the exhaustive interpretation of answers. Here it has been observed that negative answers cancel or restrict an exhaustive interpretation. Also exhaustive interpretation is standardly modelled as selecting models that are minimal with respect to some order. Another interesting fact is that many languages develop question markers out of their markers of negation.[19] Something similar has been observed for disjunction as well.

Negation is a very exciting topic that hasn't received sufficient attention, yet. But this seems to be changing. There are a number of interesting projects, also in the philosophical literature, that are concerned with the linguistic and logical properties of negations at the moment. This manuscript is just another example of this change.

References

Alonso-Ovalle, L. (2009). Counterfactuals, correlatives, and disjunction. *Linguistics and Philosophy 32*, 207–244.

Ciardelli, I., J. Groenendijk, and F. Roelofsen (2018). *Inquisitive semantics*. Oxford: Oxford University Press.

Ciardelli, I., L. Zhang, and L. Champollion (2018). Two switches in the theory of counterfactuals. A study of truth conditionality and minimal change. *Linguistics and Philosophy*.

Fine, K. (2012). Counterfactuals without possible worlds. *The Journal of Philosophy 109*(3), 221–246.

Goodman, N. (1955). *Fact, fiction and forecast*. Indianapolis/New York/Kansas City: The Bobbs-Merrill Company, Inc.

Kratzer, A. (1981a). The notional category of modality. In H.-J. Eikmeyer and H. Rieser (Eds.), *Words, worlds, and contexts*, pp. 387–394. Berlin/New York: De Gruyter.

Kratzer, A. (1981b). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic 10*, 201–216.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1981). Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic 10*(2), 217–234.

Schulz, K. (2011). If you wiggle A, then B will change. *Synthese 179*(2), 239–251.

Stalnaker, R. (1968). A theory of conditionals. In J. Cornman et al. (Eds.), *Studies in Logical Theory: essays*, pp. 98–112. Oxford: Blackwell.

Veltman, F. (1976). Prejudices, presuppositions and the theory of counterfactuals. In J. Groenendijk et al. (Eds.), *Amsterdam Papers in Formal Grammar*, Volume 1. Centrale Interfaculteit, Universiteit van Amsterdam.

[19]Thanks to Andreas Haida for pointing this out to me.