Coherence of Argumentative Dialogue Snippets: A New Method for Large Scale Evaluation with an Application to Inference Anchoring Theory

Paul Piwek
The Open University
Milton Keynes, United Kingdom
paul.piwek@open.ac.uk

Jacopo Amidei
Universitat Oberta de Catalunya
Barcelona, Spain
jamidei@uoc.edu

Svetlana Stoyanchev
Toshiba Europe Limited
Cambridge, United Kingdom
svetlana.stoyanchev@toshiba.eu

Abstract

This paper introduces a novel method for testing the components of theories of (dialogue) coherence through utterance substitution. The method is described and then applied to Inference Anchoring Theory (IAT) in a large scale experimental study with 933 dialogue snippets and 87 annotators. IAT has been used for substantial corpus annotation and practical applications. To address the aim of finding out *if and to what extent two aspects of IAT – illocutionary acts and propositional relations – contribute to dialogue coherence*, we designed an experiment for systematically comparing the coherence ratings for several variants of short debate snippets. The comparison is between original human-human debate snippets, snippets generated with an IAT-compliant algorithm and snippets produced with ablated versions of the algorithm. This allows us to systematically compare snippets that have identical underlying structures as well as IAT-deficient structures with each other. We found that propositional relations do impact on dialogue coherence (at a statistically highly significant level) whereas we found no such effect for illocutionary act expression. This result suggests that fine-grained inferential relations impact on dialogue coherence, complementing the higher-level coherence structures of, for instance, Rhetorical Structure Theory.

1 Introduction

The proper modeling of argumentation in dialogue is a long-standing challenge, raising questions about how individual and collective reasoning and argumentation are connected (Yu et al., forthcoming; Ivanova and Gubelmann, 2025). In particular, a significant question is how coherence relations in debate are connected to the propositional relations of logical reasoning, that is conflict/oppose, and inference/support. An important proposal clarifying this relation is Inference Anchoring Theory (IAT) (Reed, 2011; Reed and Budzynska, 2011; Budzynska et al., 2014). This theory aims to account for the coherence of debates and offers the tools for argument corpus development (Budzynska et al., 2014), finetuning LLMs (Wu et al., 2024), and shedding light on, for example, the role of questions in debates (Hautli-Janisz et al., 2022).

Figure 1 shows a snippet of dialogue on the Welfare state from the Moral Maze corpus (Janier, 2017) annotated with IAT. On the right-hand side, we can see four *locutions*, labeled L1 to L4. As the example shows, a locution consists of a speaker designation ('Neil' or 'Clifford') and an utterance. Each locution is anchored to a *proposition* (shown in the four left-most blue boxes) via an *Illocutionary Connection* (IC) (in the middle yellow boxes). In this case, the two first locutions' Illocutionary Connection, i.e. IC, is 'Asserting', the third is 'Assertive Questioning' and the last one is 'Asserting'. Propositions and ICs represent the propositional content and Illocutionary force, i.e. the speaker's communicative intention, from Speech Act theory (Searle, 1969).

The locutions are linked by a *transition box* signalling a locution is a response to its predecessor. Each transition between locutions is anchored, via an *Illocutionary Connection for Transition* (ICTA), to a *Propositional relation*. In our example, one transition is anchored via the 'Arguing' Illocutionary connection to the 'Inference' Propositional relation and two transitions are anchored via the 'Disagreeing' Illocutionary connection to the Conflict Propositional relation. As the example shows, 'Inference' is used when one proposition provides a reason to accept the other proposition. In contrast, 'Conflict' is used when one proposition provides a reason to not accept the other proposition.[1]

[1]IAT singles out further propositional relations, for example, *Rephrase* (when one proposition is more or less a paraphrase of the other) but we ignore them for the purpose of this paper.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14099–14119
November 4-9, 2025

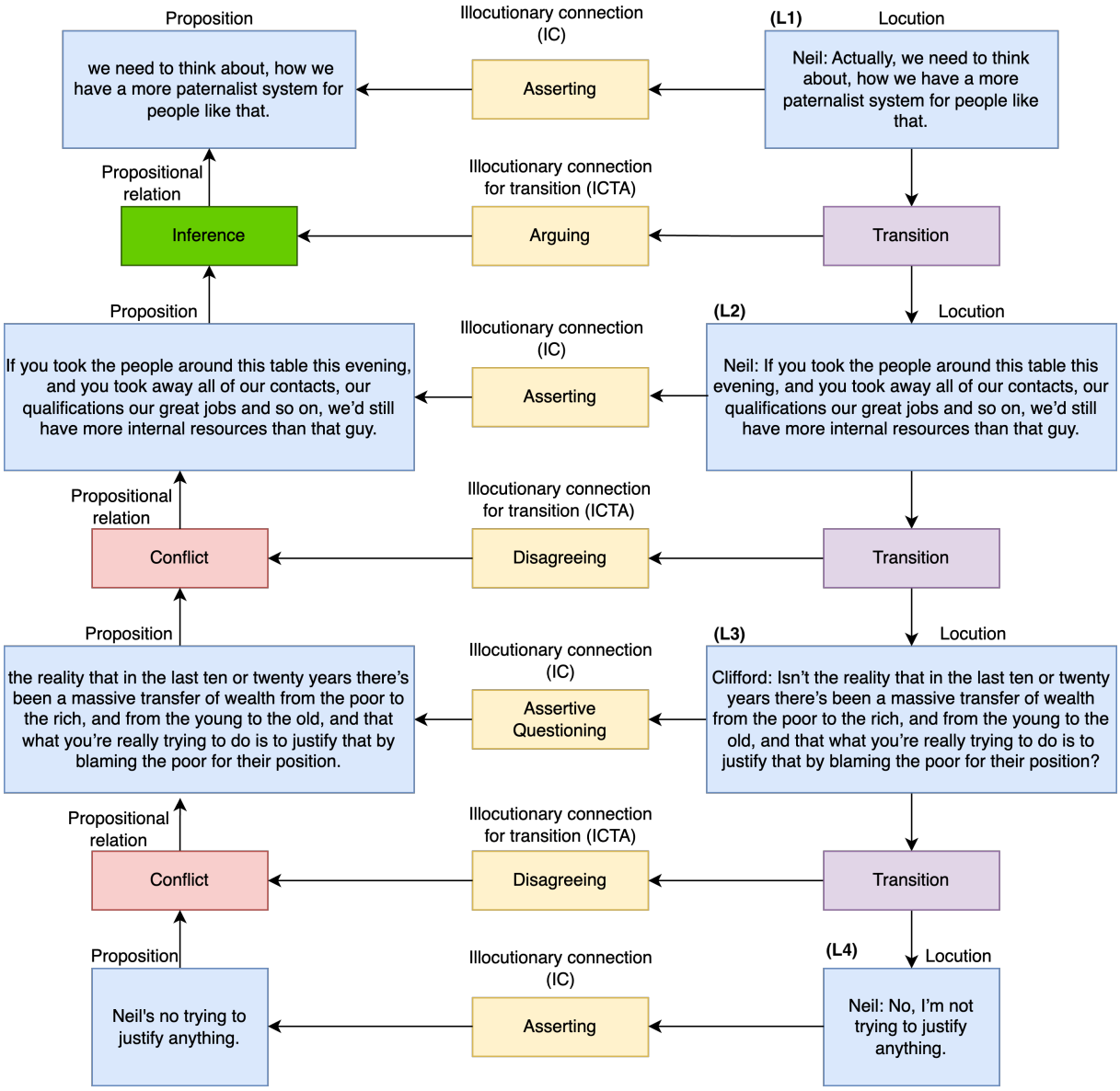

Figure 1: Example from Moral Maze Welfare State (Map6273, 2017). In this episode, Neil and Clifford are respectively a Witness and a Panellist.

The structure that IAT assigns to naturally-occurring dialogues is proposed in the tradition of theories that 'aim to account for dialogue coherence' (Budzynska et al., 2014, page 917), extending such works by using locutions as anchors for underlying propositional relations. The current paper proposes a novel way to *test if and to what extent two aspects of IAT – illocutionary acts and propositional relations – contribute to dialogue coherence*. We have developed an algorithm, Dialogue Propositional Content Replacement (DPCR), that can generate dialogue snippet variants whose structures are entirely or partially IAT-compliant. For this purpose, the algorithm replaces locutions in an IAT structure with content from an argument map (capturing inference/support and conflict/oppose relations between propositions on a given topic), but in such a way that the IAT-structure constraints (on propositional and illocutionary relations) are either satisfied, partially satisfied or not satisfied at all. For instance, replacing L1 and L2 in Figure 1 with (L1) 'Neil: I think that all humans should be vegan' and (L2) 'Neil: In the sense that a world of veganism would be a more ethical word: its morals would bring benefits to human society', using claims from a Kialo map on Veganism, results in a new dialogue snippet that satisfies the same underlying propositional and illocutionary relations as the original dialogue. The full example, including (L3) and (L4) and further examples are provided in Appendix

G. Also, the Supplementary materials for this paper include all experimental materials (https://doi.org/10.21954/ou.rd.30161896).

We obtain coherence ratings from human readers for these DPCR-generated dialogue snippets as well as for the original naturally occurring dialogue snippets and snippets generated with ablated versions of DPCR. This provides us with insight into: (1) the level of coherence of DPCR-generated dialogues in relation to the original dialogue (when the topic has been changed but the structure retained) and (2) the effect on coherence ratings as a result of the presence/absence of the two key aspects of the DPCR algorithm investigated here, i.e. propositional and illocutionary relations. Note that for this second comparison between DPCR and ablated versions, the topic is constant but the generative algorithm varies (so we are not comparing coherence across dialogue topics).

Our work makes two contributions to the study of (dialogue) coherence. Firstly, it introduces a novel method for empirically testing theories of coherence through a method of utterance substitution. Secondly, by applying our method in a large-scale empirical investigation of IAT, we show the viability of both the method, as well as the potential of certain aspects of IAT for a possible model of dialogue coherence. The method and study are offered as exemplars for further work contributing to a long-term research programme into the factors underlying coherence in discourse. As such, the current work sits firmly within Computational Linguistics, as it uses an empirical study – enabled by computational means, i.e. an algorithm for generating dialogue snippet variants – to validate certain aspects of IAT, a linguistic theory within the remit of discourse analysis and pragmatics. Identifying such aspects could lead to a possible synthesis of existing theories, as described further in the related work section.

In the next section, we describe related work. The remainder of the paper follows the standard American Psychological Association (APA) format for reporting experimental research.

2 Related work

It is notoriously difficult to evaluate discourse analysis theories and annotation schemes, such as IAT – however, we agree with the assessment of Knott (2007, Page 594), who proposes that evaluation of a theory of coherence is 'considerably more compelling as empirical support' when done by means of an application of the theory for text generation that can then be assessed against judgements from 'actual readers'. Knott (2007) argues against the evaluation of theory-based annotations directly and instead suggests that 'the theory must be assessed in relation to its predictions about independently observable phenomena in discourse', a point also made by Zaenen (2006). Knott (2007) commends Wolf and Gibson (2006) for their work on evaluating graph-based against tree-based theories by comparing the quality of summarised texts achieved by these different underlying representations. Such empirical theory evaluation seems to have, however, had little uptake and, unfortunately, may even have less chance in future where most effort goes to experimentation with large language models. The method proposed in the current paper adds an important twist to Knott and Zaenen's recommendations, by going beyond between-theory comparison, enabling (via ablation) a better understanding of individual aspects of a theory of coherence that account for coherence judgements.

Head-to-head comparison of theories of coherence, such as RST (Mann and Thompson, 1988), SDRT (Asher and Lascarides, 2003), or QUD (Ginzburg, 2012), may be challenging and even unnecessary. For instance, RST could be evaluated along the lines of our evaluation of IAT, using as a substitution-based approach. To our knowledge, the only existing work in this space, using a substitution-based approach grounded in RST, was trialled in Piwek and Stoyanchev (2011), but for monologue-to-dialogue generation rather than theory testing. A substitution-based experiment along the lines described in the current paper, but with RST, would not aim to determine whether RST is the better theory, but could be used to shed light on the elements of RST that matter. This in turn could lead to an integration of the key effective elements of RST and IAT, a possibility considered by Budzynska et al. (2014). In our view, RST and IAT are complementary in that RST has a much wider range of relations that are needed to account for a much wider range of discourse genres beyond debate, whereas IAT provides the means for dealing with fine-grained inferential relations found mostly, but not exclusively, in debate and that are beyond the scope of RST. Showing that these fine-grained relations affect judgements of coherence does *not* falsify RST – rather we propose that this is evidence of a complementary level of structure that

affects coherence.

The current research only considers short, up to four-turn, dialogue snippets, given that longer snippets typically do not have a fully connected IAT structure underpinning them. IAT works well where relations between locutions can be mapped to underlying propositional relations. Where these are absent, other factors will influence dialogue coherence that are better accounted for by theories such as RST. Thus we return to our earlier point that to fully account for dialogue coherence, in the long run, a synthesis of theories based on a thorough understanding of what makes each effective, seems the best route to success. Our experimental method and the results of its application to IAT are intended to contribute to this long-term programme.

3 Method

3.1 Materials

Coherence Rating Scale For our experiment, we developed, based on pilot studies and previous work on coherence annotation such as Cervone and Riccardi (2020), a scale from 1 to 7 for rating the coherence of argumentative dialogue snippets (with 1 = **incoherent** and 7 = **coherent**).[2] Annotaters are asked to rate a dialogue snippet as coherent if the following apply: **(a)** all sentences in the dialogue make sense by themselves and are clear (at the point in dialogue where they occur), and **(b)** all sentences in the dialogue link together well with each other so that the dialogue is clear and sensible.

Dialogue Snippet Variants To address our aim of validating IAT, the participants in the current study rated the coherence of dialogue snippets belonging to one of the following categories:

1. original naturally-occurring argumentative dialogue snippets (MORAL MAZE$_{original}$) from the Moral Maze corpus,

2. original naturally-occurring argumentative dialogue snippets, but after contextual enhancement: i.e., where there is anaphora or ellipsis, we manually expand these to make the dialogue more self-contained (MORAL MAZE$^{+context}_{original}$), since this could affect coherence ratings,

3. argumentative dialogue snippets generated by means of the Dialogue Propositional Content Replacement (DPCR) algorithm described below,

4. argumentative dialogue snippets generated by the 'No sentence templates' algorithm (DPCR$_{-templ}$). This algorithm generates new argumentative dialogues according to the same algorithm as DPCR but without applying sentence templates.

5. argumentative dialogue snippets generated by the 'Random propositional relations' algorithm (DPCR^{-rel}). This algorithm applies sentence templates corresponding with Illocutionary Connections (ICs) in locution patterns (LPs), but selects a random propositional relation rather than the relation selected according to IAT, and

6. argumentative dialogue snippets generated by the 'No sentence templates and random propositional relations' algorithm (DPCR$^{-rel}_{-templ}$). This algorithm selects a random proposition and does not apply the sentence templates to generate locutions.

Note that for our study we use four algorithms for snippet generation: the full DPCR algorithm as well as three ablated versions of this algorithm. We also have human-generated argumentative dialogues snippets (MORAL MAZE$_{original}$ and MORAL MAZE$^{+context}_{original}$). As shown in Table 1 these algorithms generated 883 dialogue snippets. Additionally, there are 50 dialogues from MORAL MAZE$_{original}$ and MORAL MAZE$^{+context}_{original}$ adding up to total of 933 snippets. The Data part of the Supplementary Materials for this paper (see https://doi.org/10.21954/ou.rd.30161896) includes the full set of dialogue snippets. Additionally, for representative examples, see Appendix G on Page 19.

Once generated, the argumentative dialogue snippets were split into batches of 13. In each batch, two snippets were repeated twice each, to be used for annotator quality control. The two repeated snippets that were presented twice at random places in the batch allowed us to assess the annotator's self-consistency. Overall, we have 71 batches of 15 dialogues (13 plus 2 repetitions) and 1 batch of 12 dialogues (10 plus 2 repetitions).

[2]The guidelines used in the experiment can be found in Appendix F on Page 18. Further detail can be found in the Data Supplementary Materials at https://doi.org/10.21954/ou.rd.30161896.

Algorithm	Brexit	Veganism	Vaccination	Total
DPCR	72	75	72	**219**
DPCR$_{-templ}$	72	75	72	**219**
DPCR^{-rel}	74	72	75	**221**
DPCR$^{-rel}_{-templ}$	75	74	75	**224**
Total	**293**	**296**	**294**	**883**

Table 1: Number of argumentative dialogue snippets generated per topic and per algorithm for Brexit, Veganism and Vaccination.

Method for Dialogue Propositional Content Replacement (DPCR) For the current work, we made use of an enhanced version of the Moral Maze MM2012c dataset (Janier, 2017): the QTMM2012c+ dataset (Amidei et al., 2021). The latter includes the following additional information: (a) each speaker is labelled with their role (one of Chair, Panellist or Witness), (b) speakers are associated with a stance towards the claim or thesis under discussion (neutral, pro and con) and (c) information on the locutions chronological order is made explicit.

The second main resource that the current work draws on are argument maps. An argument map is a tree-like structure that starts with a thesis (top claim, blue box). The thesis can be supported or attacked by *pro* (green dashed boxes) and *con* (red boxes) arguments. In turn, both pro-arguments and con-arguments can branch into subsequent arguments that support or attack them. For an example of an argument map's structure see Figure 2 in the Appendix. Argument maps and related structures such as argument graphs have been used previously to drive persuasive chatbots, see Chalaguine and Hunter (2020). The DPCR algorithm is not tied to a specific dataset of argument maps, but for our study we will be using maps with claims from Kialo.com.

In a nutshell, with DPCR we take an existing snippet of an argumentative dialogue from an argumentative dialogue corpus and replace its locutions with claims lifted from an argument map on a different topic whilst retaining the IAT dialogue structure, including propositional relations between the locutions' contents. Formally, given an argumentative dialogue snippet D consisting of the sequence of locutions $l_1, \ldots, l_n$ on topic T with:

- *locutions* as the set of possible locutions;
- $type : locutions \longrightarrow dialogue\_act\_type$; [3]
- $speaker\_role : locutions \longrightarrow roles$; [4]
- $content : locutions \longrightarrow propositions$; [5]
- $prop\_rel : propositions \times propositions \longrightarrow propositional\_relation$. [6]

Argumentative Dialogue Propositional Content Replacement (DPCR) is defined as obtaining a dialogue snippet $D' = l'_1, \ldots, l'_n$ on topic T' from a dialogue snippet $D = l_1, \ldots, l_n$ on topic T such that:

- for all $1 \leq x \leq n : type(l_x) = type(l'_x)$
- for all $1 \leq x \leq n : speaker\_role(l_x) = speaker\_role(l'_x)$
- for all $1 \leq x, y \leq n$: $prop\_rel(content(l_x), content(l_y)) = prop\_rel(content(l'_x), content(l'_y))$

This definition stipulates what counts as DPCR, i.e. replacing propositional content on topic T with content on topic T' applied to argumentative dialogue snippet D on topic T, resulting in snippet D': as we replace locutions on one topic for those on another, (a) the dialogue act types and speaker roles belonging with the replaced for locutions should remain the same and (b) where there are propositional relations between the contents of the original locutions, these should also hold between the contents

[3] For more detail on the dialogue act types used in this paper, we refer to Table 5 in Appendix B.

[4] For this paper we use the roles *chair*, *panellist* and *witness*.

[5] Propositions are represented as paraphrases of the locutions, with context-dependence removed where possible.

[6] For the purpose of this work we only distinguish two propositional relations: pro (or $inference$) and con (or $conflict$). The labels pro/con are used for propositional relations in argument maps to signify support (pro) and opposition (con) between two propositions. These correspond to the IAT propositional relations $inference$ and $conflict$. Note that by restricting our work to these two relations, the current function $prop\_rel$ is partial. For pairs of propositions where the relation between the propositions is not one of the aforementioned two relations, we assume that it maps to $\star$.

of the replacement locutions (taken from argument map). Thus we obtain a new snippet that has the same IAT structure as the original snippet, but deals with a different topic. DPCR$_{-templ}$ violates (a) by rendering almost all acts as assertions, DPCR^{-rel} violates (b) by selecting propositional relations at random, and DPCR$^{-rel}_{-templ}$ violates both (a) *and* (b). Appendix D on Page 16 contains a full description of the DPCR algorithm and its ablations, whilst the code is supplied as Supplementary Material at https://doi.org/10.21954/ou.rd.30161926.

3.2 Participants

For our experiment, we used the Amazon Mechanical Turk platform (Mturk, 2022) with 10 annotators per batch. Each annotator was paid $4, for a task of 20 minutes.[7] The annotators were Master annotators[8] from the UK and USA with as their minimum education a US Bachelor degree. We had 89 annotators who performed the task. Two of them were rejected resulting in data being used from 87 annotators. The two annotators were rejected on the basis of a test-retest setup. In each batch, the test-retest setup was based on two dialogues each being repeated once. We expected the participants to assign identical or close scores to identical (repeated) dialogues. We split the scores into three sets: $\{1, 2, 3\}$, $\{3, 4, 5\}$ and $\{5, 6, 7\}$. If the scores from a repeated dialogue were different and part of two different sets, then we consider the test failed and rejected the annotator.

3.3 Design

Our aim is to understand *which aspects, if any, of Inference Anchoring Theory (IAT) may account for dialogue coherence?* To systematically investigate this question, we introduce four testable hypotheses:

(H1) DCPR-generated snippets are at least as coherent as MORAL MAZE snippets.

(H2) DCPR-generated snippets are more coherent than DPCR^{-rel} snippets.

(H3) DCPR-generated snippets are more coherent than DCPR$_{-templ}$ snippets.

(H4) DCPR-generated snippets are more coherent than DPCR$^{-rel}_{-templ}$ snippets.

The first hypothesis (H1) checks that the level of coherence of IAT-structured generated dialogue snippets is at least at the same level as that of natural dialogues. This comparison aims to ascertain that the original and IAT-compliant dialogues have a similar/comparable level of coherence. This hypothesis is subservient to the purpose of the remaining hypotheses (H2-H4), which is the comparison of the IAT-compliant dialogues with non-compliant/ablated ones on the same dialogue topics.

The hypotheses H2-H4 compare snippets that are fully IAT-compliant with those that are only partially or not at all compliant. Together, these hypotheses tests to what extent the structures posited by IAT allow us to create dialogue snippets that, on the one hand, are comparable in coherence with snippets from naturally-occuring dialogue and, on the other hand, are superior in coherence when compared with dialogue snippets that at best only partially conform with with IAT dialogue structure.

3.4 Procedure

Annotators judged one debate snippet at a time. Snippets were grouped into batches, as described above, where the order was randomised per participant (to avoid ordering effects). Annotators could annotate more than one batch, but never the same batch twice.

4 Results

Annotator reliability and coherence ratings Table 2 reports the Inter Annotator Agreement (IAA) value measured with two different metrics, to provide a good overview of the data reliability.[9] The values in Table 2 are based on the IAA for each of 72 batches.

In our experiment, the lower categories 1–4 are used much less than the higher categories 5–7. This makes our annotation unbalanced towards the lower categories. Under such conditions, chance-corrected coefficients such Krippendorff's α (Krippendorff, 1980), Fleiss's κ (Fleiss, 1971) and Cohen's κ (Cohen, 1960) are subject to the prevalence paradox (Artstein and Poesio, 2008) and sub-optimal. For this reason, we decided to report IAA

[7] With the task taking up to 20 minutes at £3.38 (based on exchange rate at the time), this amounts to remuneration at £10.14/hour. When we carried out the experiment, in July 2022, the minimum wage in the UK was £9.50/hour.

[8] Master Workers are a top Worker of the MTurk marketplace. For more details see Mturk FAQs (2022).

[9] All the criteria were measured by the use of *irrCAC* library provided by the *R* software (Gwet, 2014b). More specifically we used the functions pa.coeff.raw() and gwet.ac1.raw(), all with ordinal weight.

Value	%	*AC2*
Mean	0.82	0.39
Max	0.92	0.80
Min	0.77	0.21
Median	0.84	0.49
Variance	0.0008	0.01

Table 2: Value of Inter Annotator Agreement measured among batches. Where % is the Percent Agreement and *AC2* is the Gwet AC2 coefficient.

based on the Gwet *AC2* coefficient (Gwet, 2014a) which is deemed to be more robust.

To interpret the IAA values we used the Landis and Koch (1977) benchmark scale as revised and adjusted by Gwet (2014a).[10] Based on this analysis we got a level of agreement equal to or higher than *fair* for 83% of the batches. More precisely, 3% of the bathes reached a *substantial* level of agreement, 30% of the bathes reached a *moderate* level of agreement and 50% of the batches reached a *fair* level of agreement. Finally, a *slight* level of agreement was reached for 17% of the batches. Judging the coherence of a dialogue is not straightforward. Many factors can impact dialogue coherence, and make a dialogue more or less coherent. This made the task of judging dialogue coherence a subjective one. Accordingly, we consider the agreement reached in our study a satisfactory level of agreement.

Hypotheses Table 3 shows the results of our empirical evaluation. Table 20 in Appendix H breaks these results further down by comparing the DPCR-generated variants per topic. The results from the current Table 3 are reproduced, reassuring us that our results are not topic-dependent. We proceed with describing our results in terms of the hypotheses from Section 3.3.

(H1) *DCPR-generated snippets are at least as coherent as* MORAL MAZE *snippets.* This hypothesis is confirmed: DCPR-generated snippets are not just as coherent as MORAL MAZE snippets but even more coherent (according to the raters): coherence of DCPR is higher than MORAL MAZE$_{original}$ and MORAL MAZE$^{+context}_{original}$ (5.88 versus 4.9 and 5.16, $P \leq 0.001$).

[10] Also in this case, to interpret the IAA values, we used the *irrCAC* library provided by the *R* software (Gwet, 2014b). More specifically we used the functions landis.koch.bf().

Algorithm	Med. Coh.	Mean Coh.
DPCR$_{-templ}$	7	5.99
DPCR	6	5.88
MORAL M $^{+context}_{original}$	5	5.16***
MORAL M$_{original}$	5	4.9***
DPCR$^{-rel}_{-templ}$	5	4.66***
DPCR^{-rel}	5	4.46***

Table 3: Experiment results. **Med/Mean Coh. Score** is the median/mean of the coherence scores given to an algorithm. *** indicates that the difference between the algorithm in this row and DPCR, measured by the Student's t-test, is highly significant (at $P \leq 0.001$).

(H2) *DCPR-generated snippets are more coherent than DPCR^{-rel} snippets.* This hypothesis is also confirmed (5.88 versus 4.46, $P \leq 0.001$).

(H3) *DCPR-generated snippets are more coherent than DCPR$_{-templ}$ snippets.* This hypothesis could not be confirmed. There is no statistically significant difference between DCPR-generated and DCPR$_{-templ}$ snippets (5.88 versus 5.99).

(H4) *DCPR-generated snippets are more coherent than DPCR$^{-rel}_{-templ}$ snippets.* This hypothesis is also confirmed (5.88 versus 4.46, $P \leq 0.001$).

Table 4 shows the mean scores depending on the number of turns per dialogue. The table suggests that the perceived dialogue coherence is impacted by the number of turns. For the DPCR and DPCR$_{-templ}$-generated snippets, as there are more turns, the score decreases gradually. In contrast, for MORAL MAZE$^{+context}_{original}$ and MORAL MAZE$_{original}$ the score increases as the number of turns increases. Overall, the trend is that as turn number increases, the diffence between coherence levels of, on the one hand, the DPCR and DPCR$_{-templ}$-generated snippets and, on the other hand, the MORAL MAZE$^{+context}_{original}$ and MORAL MAZE$_{original}$ disappears. In contrast, for the ablated versions DPCR^{-rel} and DPCR$^{-rel}_{-templ}$, as dialogue length increases coherence decreases.

5 Discussion

Inference Anchoring Theory (IAT) is a widely used theory that presents an appealing perspective on how dialogue coherence is underwritten by an underlying structure involving illocutionary acts and logical relations of conflict and inference between propositional contents. The current research aimed to examine which aspects of IAT might account

Algorithms	2 Tns	3 Tns	4 Tns
DPCR$_{-templ}$	6.27	5.84	5.5
DPCR	6.13	5.63	5.41
Moral maze$^{+context}_{original}$	5.42	4.56	5.35
Moral maze$_{original}$	4.89	4.34	5.53
DPCR$^{-rel}_{-templ}$	4.78	4.56	4.44
DPCR^{-rel}	4.64	4.28	4.28

Table 4: Average score per number of turns (Tns).

for dialogue coherence. Using our novel experimental method, we collected the, to our knowledge, first empirical evidence that some of the underlying structure IAT assigns to debates accounts, at least partially, for the coherence of those debates. To address our overall aim of finding out *if and to what extent two aspects of IAT (illocutionary acts and propositional relations) contribute to dialogue coherence*, we tested four hypotheses.

Three of our hypotheses – (H1), (H2) and (H4) – were confirmed: We saw that IAT-generated snippets were at least as coherent as naturally-occurring dialogue snippets from the Moral Maze corpus (H1). We also saw that dialogue snippets whose underlying propositional relations were selected at random (which is the case for both DPCR^{-rel} and DPCR$^{-rel}_{-templ}$), rather than driven by IAT, are judged to have lower coherence than dialogue snippets that conform with the propositional relations mandated by IAT (H2 and H4). These results suggest that the propositional relations posited by IAT and their anchoring in dialogue is a factor that influences dialogue coherence. The other factor we considered, illocutionary acts, was, however, not confirmed by our study. More specifically, (H3) regarding DPCR$_{-templ}$ was not confirmed. In other words, judgements of coherence were not affected negatively by whether the illocutionary force (question version assertion) of the dialogue acts was reflected in their surface verbalisation (specifically through interrogative form for questions). This result may have several reasons: (1) inadequacy of the handcrafted generation templates for illocutionary force, (2) the fact that even in the DPCR$_{-templ}$ condition for those utterances without a propositional relation, illocutionary force was always expressed (as we explain below it was impossible not to do so) and (3) the ability of readers to infer implicit force. This final point is an empirical fact. Though DPCR$_{-templ}$ does not convert argument map claims into questions (i.e. interrogative sentences) where the Illocutionary Connection requires this, in dialogue, whether a locution with a declarative sentence type is intended as a question can usually be inferred from the communicative context (Beun, 1990). In our argumentative dialogue set-up, involving discussions between a panellists and witnesses about contentious topics, a natural interpretation of locutions consisting of a declarative sentence is as raising questions for discussion by the other party – this is also in line with the more general idea that assertions can can initiate issues, i.e. introduce questions for discussion, which then become part of the QUD, i.e. questions under discussion (Ginzburg, 2012). In the limitation section, we will return to points (1) and (2), providing examples that illustrate the concrete limitations of our approach in this respect.

An at first sight surprising result was that the DPCR and DPCR$_{-templ}$ algorithm-generated snippets were judged as *more* coherent than the original dialogue snippets. A possible reason for this is that the claims that the algorithm takes from the Kialo maps are generally well-written and self-contained. In contrast, some of the original spoken language locutions in the Moral Maze snippets are less self-contained and context-dependent. In our experimental design we compensated for this by creating versions of the Moral Maze snippets with added context. This helped somewhat, with MORAL MAZE$^{+context}_{original}$ dialogues rated slightly higher than MORAL MAZE$_{original}$, but still not at the level of DPCR and DPCR$_{-templ}$. In any event, our main comparison was between the (ablated) versions of DPCR, allowing us to determine whether propositional relations and/or illuctorionary act (expression) play a role in dialogue coherence ratings. The fact that DCPR dialogues received coherence ratings comparable to those of the original dialogues reassures us that comparing DCPR dialogue ratings with ablated version ratings is warranted.

Another interesting finding is that participants mostly use the upper end of the rating scale (4–7) for coherence ratings. This may be explained by the general tendency of interpreters to expend effort to make sense of utterances even if, on the face of it, they do not make sense. An insight that has been widely discussed following on from the introduction of the notion of implicatures (Grice, 1975) (i.e. additional inferences that addressees make to explain away any apparent lack of cooperativeness of speaker contributions). This does, however, not undercut the idea central to the current project, i.e.

that there are degrees of coherence.

Apart from validating the role of IAT propositional relations in dialogue coherence, the current work contributes to the computational study of argumentative dialogue by offering the DPCR algorithm (full and ablated versions) and its implementation for use by the research community as well as the corpus of 933 generated and natural-occurring dialogue snippets together with their coherence ratings, with each snippet rated by 10 annotators out of a group of 87 annotators. This dataset (see https://doi.org/10.21954/ou.rd.30161896) may also prove valuable for assessing to what extent LLMs as judges agree with human dialogue coherence judgements in our dataset, especially since the dataset was collected just before the release of ChatGPT (and thus not subject to the current risk of raters on crowdsourcing platforms such as MTurk making use of LLMs).

6 Limitations

Contra our hypothesis (H3), the DPCR$_{-templ}$ algorithm-generated snippets were rated as highly in terms of coherence as DPCR-generated snippets. We had expected that by switching off the template generation, the coherence would detioriate. The idea behind the template generation was to convert propositions (stated as assertions) into the correct dialogue acts, i.e. the act type observed in the original naturally-occuring snippet from which the generated snipped was derived via DPCR.

A qualitative analysis of the dialogue snippets generated with DPCR$_{-templ}$ and DPCR revealed coherence score differences where the snippets begin with a speaker that uttered a questioning followed by an asserting. This suggests that there is scope to improve the relevant sentence templates for this situation. For example:

> ***Assertive Questioning:*** *Do you believe that the UK should remain in the EU if a hard Brexit is the only alternative option?*

followed by:

> ***Asserting:*** *In other words I think that by remaining in the EU, the UK would be able to operate broadly as before but with clear caveats regarding some issues that concern its citizens.*

In cases like this, a dialogue generated with the DPCR$_{-templ}$ algorithm can result in dialogue that is perceived as more coherent. For example:

> *The UK should remain in the EU if a hard Brexit is the only alternative option.*

followed by:

> *By remaining in the EU, the UK would be able to operate broadly as before but with clear caveats regarding some issues that concern its citizens.*

As illustrated above, the DPCR$_{-templ}$ algorithm does not make use of the Illocutionary Connections (ICs) to rephrase the argument map claims that are used. It will *not* convert the argument map claim to the sentence type associated with the IC and instead always use the declarative sentence from the argument map verbatim. However, for those locutions where there is no argument map claim involved, such as various forms of challenging, DPCR$_{-templ}$ does use the relevant canned text: for example a *Pure challenging* IC is realised as one of the following 'Why is that?' or 'Why?'. Similarly, there is canned text for *Assertive* and *Rhetorical Challenging*. This means that DPCR$_{-templ}$ will not just yield a sequence of assertions: for all the aforementioned ICs involving challenging, variety is introduced through the canned text associated with these ICs. All in all this means that DPCR$_{-templ}$ is at least partially IAT-compliant after all.

As part of our qualitative analysis we also observed that the dialogue generated with DPCR does look more like natural dialogue, than the ones generated with the DPCR$_{-templ}$ algorithm. For an example of the contrast between the two types of dialogue that are generated, see Appendix G starting on Page 19: Tables 18 and 19. Note the difference between the DPCR dialogue, which involves for example questions and hedges and DPCR$_{-templ}$ dialogue, which is a simple sequence of assertions. It may be that naturalness needs to be considered separately from coherence, which was the focus of this paper. Whereas we found evidence for the relation between propositional relation choice and coherence, this relation does not seem to be as strong or existent between dialogue act type choice and coherence. Further research is needed to establish whether the latter relation is more closely associated with dialogue naturalness.

In our Related Work section we discuss our decision to limit the scope of the current work to short

dialogue snippets rather than full-length dialogues. In our view, widening the scope to full-length dialogues requires a synthesis of theories – such as IAT with RST, SDRT or QUD, or elements of theories going further back presented in the work of e.g. Polanyi and Scha (1984) or Grosz and Sidner (1986) – based on a thorough understanding of what makes each effective. Here, we would like to emphasize that our experimental method and the results of its application to IAT are intended as a contribution to such a long-term research programme.

We consider the current study large scale, with almost 1000 dialogue snippets judged by 87 annotators. Of course, size is relative and compared to datasets and annotations undertaken by commercially-driven labs (e.g. to train LLMs), our dataset is comparatively small. And yet, for theory-driven empirical work, this study is of a significant size. It is also worth noting that with this paper we are making all data (https://doi.org/10.21954/ou.rd.30161896) and code (https://doi.org/10.21954/ou.rd.30161926) available.

Acknowledgements

This research was carried out as part of the Opening Up Minds project, funded by the UK Engineering and Physical Sciences Research Council under grant EP/T024666/1. We would like to thank the project team members of the linked grants EP/T023414/1 and EP/T023554/1 for discussion of this work: Lotty Brand, Youmna Farag, Tom Stafford and Andreas Vlachos. We would also like to thank the ARR reviewers for their helpful feedback.

Ethics approval

The research protocol for the current research project was given a favourable opinion by the The Open University's Research Ethics Committee.

References

Jacopo Amidei, Paul Piwek, and Svetlana Stoyanchev. 2021. QTMM2012c+: A Queryable Empirically-Grounded Resource of Dialogue with Argumentation. In *5th Workshop on Advances in Argumentation in Artificial Intelligence*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, United States.

Robbert-Jan Beun. 1990. The recognition of dutch declarative questions. *Journal of Pragmatics*, 14(1):39–56.

Kasia Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yakorska. 2014. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 917–924, Reykjavik, Iceland. European Language Resources Association (ELRA).

Centre for Argument Technology. 2023 manuscript. A Quick Start Guide to Inference Anchoring Theory (IAT).

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this dialogue coherent? learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174, 1st virtual meeting. Association for Computational Linguistics.

Lisa A Chalaguine and Anthony Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *Computational Models of Argument*, pages 9–20. IOS Press.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jonathan Ginzburg. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Kilem L. Gwet. 2014a. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Kilem L. Gwet. 2014b. irrCAC library home page. https://rdrr.io/cran/irrCAC/. [Online; accessed 2022].

Annette Hautli-Janisz, Katarzyna Budzynska, Conor McKillop, Brian Plüss, Valentin Gold, and Chris Reed. 2022. Questions in argumentative dialogue. *Journal of Pragmatics*, 188:56–79.

Rositsa V Ivanova and Reto Gubelmann. 2025. The shift from logic to dialectic in argumentation theory: Implications for computational argument quality

assessment. In *Procs of the 31st Int Conf on Computational Linguistics*, pages 4789–4802, Abu Dhabi, UAE. Association for Computational Linguistics.

Mathilde Janier. 2017. *Dialogical dynamics and argumentative structures in dispute mediation discourse*. Ph.D. thesis, University of Dundee.

Alistair Knott. 2007. Book Reviews: Coherence in Natural Language: Data Stuctures and Applications, by Florian Wolf and Edward Gibson. *Computational Linguistics*, 33(4):591–595.

Klaus Krippendorff. 1980. *Content analysis; an introduction to its methodology*. A Sage Publications, Beverly Hills, CA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

MM2012c Map6273. 2017. Moral Maze map 6273. http://ova.arg-tech.org/analyse.php?url=local&plus=true&aifdb=6273&akey=49adc508e74cadf6633d666f9644000e. [Online; accessed 2022].

Mturk. 2022. Mturk home page. https://www.mturk.com/. [Online; accessed 2022].

Mturk FAQs. 2022. Mturk FAQs home page. https://www.mturk.com/worker/help. [Online; accessed 2022].

Paul Piwek and Svetlana Stoyanchev. 2011. Data-oriented monologue-to-dialogue generation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 242–247, Portland, Oregon, USA. Association for Computational Linguistics.

Livia Polanyi and Remko Scha. 1984. A syntactic approach to discourse semantics. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 413–419, Stanford, California, USA. Association for Computational Linguistics.

Chris Reed. 2011. Implicit speech acts are ubiquitous. Why? They join the dots. In *Argument cultures: Proceedings of the 8th international conference of the Ontario Society for the Study of Argumentation.*

Chris Reed and Katarzyna Budzynska. 2011. How dialogues create arguments. In *Proceedings of the 7th conference on argumentation of the International Society for the Study of Argumentation*, pages 1633–1645.

John R. Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Kialo Veganism example. 2022. All humans should be vegan. https://www.kialo.com/all-humans-should-be-vegan-2762?path=2762.0~2762.1. [Online; accessed 2022].

Florian Wolf and Edward Gibson. 2006. *Coherence in natural language: data structures and applications*. MIT Press.

Yuetong Wu, Yukai Zhou, Baixuan Xu, Weiqi Wang, and Yangqiu Song. 2024. KnowComp at DialAM-2024: Fine-tuning Pre-trained Language Models for Dialogical Argument Mining with Inference Anchoring Theory. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 103–109, Bangkok, Thailand. Association for Computational Linguistics.

Liuwen Yu, Réka Markovic, and Leendert Van der Torre. forthcoming. Thirteen Challenges in Formal and Computational Argumentation. In *Handbbook of Formal Argumentation Vol. 3. / Journal of Applied Logics*.

Annie Zaenen. 2006. Last words: Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.

APPENDIX

A Example of an argument map

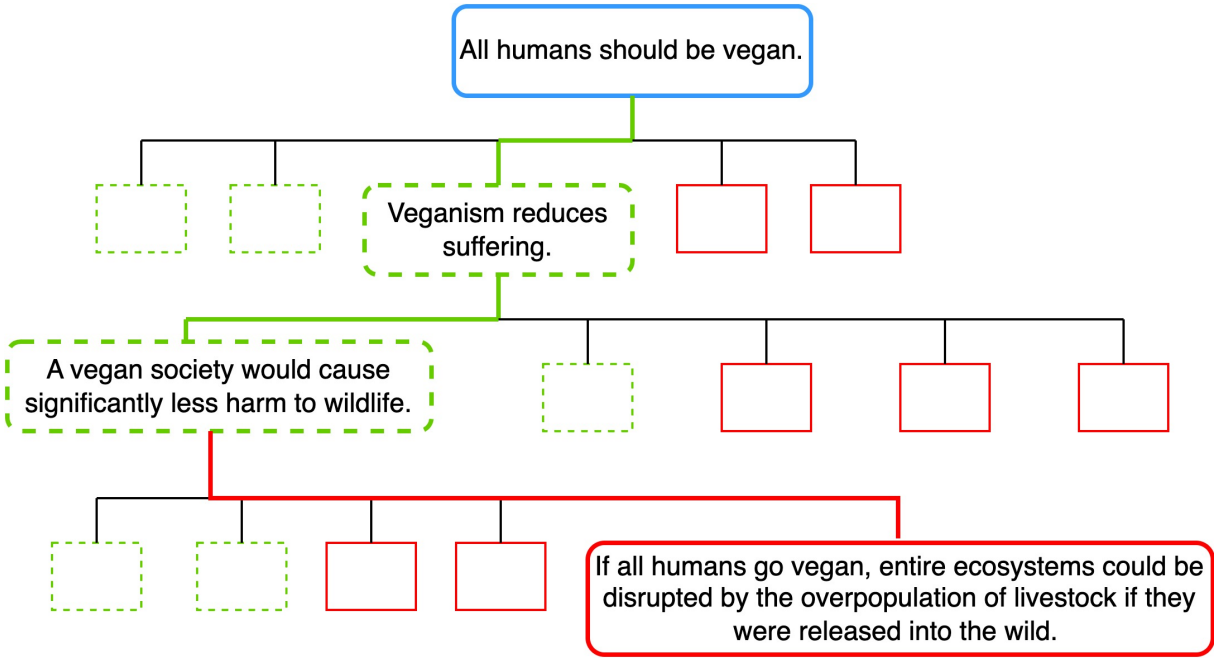

Figure 2: Example of argument map about veganism, with claims based on (Veganism example, 2022).

B IAT Dialogue Act Types

Illoctionary Connection (IC)
Questioning
Rhetorical Questioning
Assertive Questioning
Pure Questioning
Challenging
Rhetorical Challenging
Assertive Challenging
Pure Challenging
Others
Asserting
Popular Conceding
Yes
No

Table 5: Types of Dialogue Acts, referred to as Illocutionary Connections (ICs) in Inference Anchoring Theory (IAT). Detailed descriptions of these acts can be found in Centre for Argument Technology (2023 manuscript)

C Student's t-test statistics

The Student´s t-test was used for comparing the DPCR algorithm with the other algorithms.[11] The measure was performed on coherence scores associated with the dialogue generated by each algorithm. In the same fashion, we performed the Student's t-test per topic. In this case, we focus on the coherence scores associated with dialogue snippets grouped by topic.

Table 6 reports statistics related to the Student´s t-test. Similarly, Tables 7, 9, 8 report statistics related to the Student´s t-test respectively for the case of Brexit, vaccination and veganism. Table 10 and Table 11 report respectively the standard deviation of the coherence scores measured for each algorithm and the standard deviation of the coherence scores measured for each algorithm per topic. Table 12 and Table 13 report respectively the variance of the mean coherence scores and the variance of the mean coherence scores per topic.

Algorithms	**t-test score**	**p-value**	**Degree of freedom**
DPCR / MORAL MAZE$_{original}$	5.77	1.442593054380911e-06	35.58
DPCR / MORAL MAZE$^{+context}_{original}$	4.53	8.436201745497882e-05	30.48
DPCR / DPCR$_{-templ}$	-1.88	0.06	504.56
DPCR / DPCR^{-rel}	19.50	2.22629836154952e-63	464.35
DPCR / DPCR$^{-rel}_{-templ}$	15.99	3.2683704258661824e-46	464.55

Table 6: Student´s t-test statistics.

Algorithms	**t-test score**	**p-value**	**Degree of freedom**
DPCR / DPCR$_{-templ}$	-0.44	0.65	159.47
DPCR / DPCR^{-rel}	10.69	1.2646138185639769e-20	154.46
DPCR / DPCR$^{-rel}_{-templ}$	9.49	3.4813207445337696e-17	157.86

Table 7: Student´s t-test statistics for the topic Brexit.

Algorithms	**t-test score**	**p-value**	**Degree of freedom**
DPCR / DPCR$_{-templ}$	-1.68	0.09	169
DPCR / DPCR^{-rel}	11.54	5.687575374991213e-23	152.14
DPCR / DPCR$^{-rel}_{-templ}$	8.88	1.3842848206119827e-15	156.84

Table 8: Student´s t-test statistics for the topic veganism.

[11] We used the function $stats.ttest\_ind()$ provided by the python library SciPy for the Student´s t-test, setting the variable $equal\_var = False$.

Algorithms	t-test score	p-value	Degree of freedom
DPCR / DPCR$_{-templ}$	-1.14	0.25	165.53
DPCR / DPCR^{-rel}	11.71	2.0820200193845604e-23	154.47
DPCR / DPCR$^{-rel}_{-templ}$	9.29	2.15709639836805e-16	144.41

Table 9: Student´s t-test statistics for the topic vaccination.

Algorithms	Standard deviation
DPCR$_{-templ}$	1.38
DPCR	1.42
MORAL MAZE$^{+context}_{original}$	1.79
MORAL MAZE$_{original}$	1.83
DPCR$^{-rel}_{-templ}$	1.93
DPCR^{-rel}	1.97

Table 10: Standard deviation of the coherence scores.

Algorithms	Brexit	Veganism	Vaccination
DPCR$_{-templ}$	1.42	1.35	1.34
DPCR	1.44	1.44	1.36
DPCR$^{-rel}_{-templ}$	1.95	1.91	1.92
DPCR^{-rel}	1.97	1.98	1.96

Table 11: Standard deviation of the coherence scores per topic.

Algorithms	Variance
DPCR$_{-templ}$	0.53
DPCR	0.47
MORAL MAZE$^{+context}_{original}$	0.62
MORAL MAZE$_{original}$	0.85
DPCR$^{-rel}_{-templ}$	0.99
DPCR^{-rel}	0.83

Table 12: Variance of the mean coherence scores.

Algorithms	Brexit	Veganism	Vaccination
DPCR$_{-templ}$	0.70	0.54	0.37
DPCR	0.45	0.55	0.43
DPCR$^{-rel}_{-templ}$	0.97	0.95	1.07
DPCR^{-rel}	0.87	0.89	0.71

Table 13: Variance of the mean coherence scores per topic.

Algorithms	Mann-Whitney U score	p-value
DPCR / MORAL MAZE$_{original}$	1424.5	1.4637139234284339e-09
DPCR / MORAL MAZE$^{+context}_{original}$	1445.5	5.196092755480786e-07
DPCR / DPCR$_{-templ}$	27898	0.0051
DPCR / DPCR^{-rel}	7118	8.28841006753334e-51
DPCR / DPCR$^{-rel}_{-templ}$	10150	1.0811689499536537e-41

Table 14: Mann-Whitney U test statistics. For measuring the Mann-Whitney U test we used the function $stats.mannwhitneyu()$ provide by the python library SciPy.

Algorithms	**Mann-Whitney U score**	**p-value**
DPCR / DPCR$_{-templ}$	3100	0.13
DPCR / DPCR^{-rel}	832.5	1.0340781888834896e-17
DPCR / DPCR$^{-rel}_{-templ}$	1117.0	2.578864793608788e-15

Table 15: Mann-Whitney U test statistics for the topic Brexit.

Algorithms	**Mann-Whitney U score**	**p-value**
DPCR / DPCR$_{-templ}$	2990.5	0.019
DPCR / DPCR^{-rel}	736.5	6.949207593279784e-19
DPCR / DPCR$^{-rel}_{-templ}$	1145.5	4.394640292547241e-15

Table 16: Mann-Whitney U test statistics for the topic veganism.

Algorithms	**Mann-Whitney U score**	**p-value**
DPCR / DPCR$_{-templ}$	3180	0.08
DPCR / DPCR^{-rel}	766.5	2.4894342072291853e-18
DPCR / DPCR$^{-rel}_{-templ}$	1104.5	4.427842296383658e-15

Table 17: Mann-Whitney U test statistics for the topic vaccination.

D The DPCR Algorithm

To describe the DPCR Algorithm, we need to first define a number of lists, sets and functions:

- Lists and sets:
  - Arg_{map} is the list of all the claims that make up an argument map.
  - $\text{Sentence}_{templates} = \{IC_1, \ldots, IC_m\}$, where each IC_i (for $i = 1, \ldots, m$) is a set of sentence templates for an Illocutionary Connection (IC).
  - $\text{Final}_{dialogue}$ is a list of locutions that make up the generated dialogue.
- Functions:
  - $Random$ is a function that takes a set as an input and returns a random element of the input set as an output.
  - $GenerateLocution$ is a function that takes a speaker role, a claim (from an argument map) and a sentence template as input and combines them into a locution that is returned as the output. The function removes (if present) any word repetition between the argument map claim and the sentence template. Finally, the function adds a question mark to the output sentence when the IC involved is a questioning.
  - $ChildClaim_{pro;con}$ is a function that takes an argument map, a propositional relation (pro or con), and a parent claim as input and gives as output a claim that stands in the pro or con relation (of the input) to the parent claim in the argument map. If such a proposition does not exist, the function gives the string 'FinishedBranch' as an output.
  - $Remove$ is a function that takes an argument map and a claim and removes that claim from the argument map.
  - Add is a function that takes a list (L) a locution (Loc) and an index (i) and adds the locution Loc into L at the index i.

Furthermore, an argumentative dialogue pattern (ADP) is a sequence of locution patterns (LP). An LP is defined in terms of the following components:

$$(LP) \quad [Speaker\ Role, Stance, Prop.\ Relations\ List, Illocutionary\ Connection, L_{ID}]$$

where:

- *Speaker Role* is one of the following: Chair, Witness, Panellist. It represents the role of the speaker.
- *Stance* is one of: Pro, Con, Neutral. It represents the stance of the speaker towards the main thesis/claim.
- *Prop. Relations List* is a list such that:
  - *Prop. Relations List* = [NA]. In this case the Locution Pattern (LP) expresses the claim at the root of the argument map, the map's main thesis. Or:
  - *Prop. Relations List* = [Propositional Relation, $ParentID$], where *Propositional Relation* can be one among Con, Pro, Disagreeing and Agreeing and $ParentID$ is the parent claim connected via the Propositional Relation.[12]
- *Illocutionary Connection (IC)* is the illocutionary connection that is linked to the LP's sentence (L_{ID}).
- L_{ID} is a unique identifier/label for the LP.

We use $MainClaim$ to stand for the main claim/thesis of an argument map – this is the claim that sits at the root of the map: it can have child claims (pro and con claims), but no parent claims. Finally, given a list L, with $L[i]$ we mean the element at index i of L.

[12]Note that 'Agreeing' and 'Disagreeing' are strictly speaking not relations between propositions. Rather they have to be understood either affirmation or denial of the proposition in question. In contrast, 'Pro' and 'Con' represent a relation between *two* propositions: one being in support or contradiction with the other.

Algorithm 1: DPCR Algorithm

Input: ADP, Arg$_{map}$, Sentence$_{templates}$
Output: Final$_{dialogue}$

for $LP \in ADP$ **do**
  $Prop_{rel} = LP[PropRelList[0]]$;
  $role$ = $LP[Speaker\_Role]$;
  $SelectedTempl = Random(Sentence_{templates}[LP[IllocutionaryConnection]])$;
  if $Prop_{rel} = NA$ **then**
    $C_x = MainClaim$
  else
    $Parent_{claim} = LP[PropRelList[1]]$;
    $C_x = ChildClaim_{pro;con}(Arg_{map}, Prop_{rel}, Parent_{claim})$;
    if $C_x = FinishedBranch$ **then**
      Final$_{dialogue}$ = empty list ; `/* if the end of a branch reached, discard the dialogue */`
      End the algorithm;
    else
      $Remove(Arg_{map}, C_x)$; `/* avoids sentence repetition */`
    end
  end
  $Loc_x = GenerateLocution(role, SelectedTempl, C_x)$;
  $Add(Final_{dialogue}, Loc_x, x)$
end

E Kialo maps used for generation

Kialo Terms of Service permit "crawling" and "use our export functionality to download debates for private use." (https://www.kialo.com/terms) Accordingly, we downloaded a set of debates for our experiments, but cannot redistribute the maps themselves with this paper. However, we can share the names of the specific maps that we used so other researchers can download these maps for their use in accordance with the aforementioned Terms of Service:

Brexit

1. Brexit: was it a good choice for the UK?
2. Should the UK remain in the EU if the only alternative is a hard Brexit?
3. Should the United Kingdom Remain A Member of the European Union?

Veganism

1. All humans should be vegan.
2. Is veganism a natural right?
3. Should people go vegan if they can?
4. The ethics of eating animals: Is eating meat wrong?

Vaccination

1. Do we need a vaccine to fight the Covid 19 pandemic?
2. Is Covid 19 more dangerous than regular flu viruses?
3. Is herd immunity for Covid 19 achievable?
4. It should be compulsory for those working with the elderly to take a Covid 19 vaccine.
5. Should Covid 19 vaccines be mandatory?
6. Should vaccinations be mandatory?

F Annotator Guidelines

Thank you for participating in this study. You are free to stop participating in the study at any time you want.

In the task, you will be presented with an argumentative dialogue. You will then be asked to carefully read it and judge it. In total, you will be presented with 15 dialogues.

Before starting the task, please read the following guidelines carefully. Do also feel free to refer back to these guidelines at any time during the annotation process. Indeed, we encourage you to read these guidelines anytime you have some doubts. The task should take you about fifteen-twenty minutes.

You will be asked to judge the coherence of the dialogue on a scale from 1 to 7 (1 being **incoherent** and 7 being **coherent**).

For this study, please try to use the following definition of coherence:

A dialogue is coherent if the following apply:

1) all sentences in the dialogue make sense by themselves and are clear (at the point in dialogue where they occur);

2) all sentences in the dialogue link together well with each other so that the dialogue is clear and sensible.

As a rule of thumb, if you believe that no sentences in the dialogue come out of the blue and the sentences in the dialogue are linked together well, then please rank the dialogue coherence as 7. Conversely, if you believe that all sentences in the dialogue come as out of the blue and the sentences in the dialogue are not linked together well, then please rank the dialogue coherence as 1. In the other cases, pick a number between 2 to 6 that you believe describes the level of coherence of that dialogue.

Please note, if the speakers (who will be labelled as Chair, Witness and Panellist) are in disagreement with each other, this does not mean that the dialogue is incoherent. Speakers can have a coherent dialogue although there is a disagreement between them. Remember, a dialogue is coherent if all its sentences are clear, make sense and go well with each other.

Judging the coherence of a dialogue is not straightforward. Many factors can impact dialogue coherence, and make a dialogue more or less coherent. We ask you to judge the coherence of a dialogue based on a seven-point scale which ranges from incoherent to coherent. Please try to be consistent with your judgements throughout the evaluation.

Finally, when judging the coherence of a dialogue please do not be influenced by whether you agree with the arguments in the dialogue. Remember, coherence is independent of what you think about the topic under discussion.

G Examples of original and generated dialogues

Dialogue source / Algorithm	Example of a generated dialogue
Moral maze$_{original}$	**Witness:** (L1) Actually, we need to think about, how we have a more paternalist system for people like that. (L2) If you took the people around this table this evening, and you took away all of our contacts, our qualifications our great jobs and so on, we'd still have more internal resources than that guy. **Panellist:** (L3) Isn't the reality that in the last ten or twenty years there's been a massive transfer of wealth from the poor to the rich, and from the young to the old, and that what you're really trying to do is to justify that by blaming the poor for their position? **Witness:** (L4) No, I'm not trying to justify anything.
Moral maze$^{+context}_{original}$	**Witness:** (L1) Actually, we need to think about, how we have a more paternalist system for people like, for example, a young guy that lost his job at Tesco, mainly because he wasn't turning up to work on time, which is mainly because he was smoking a lot of spliff and he was basically very disorganised. (L2) If you took the people around this table this evening, and you took away all of our contacts, our qualifications our great jobs and so on, we'd still have more internal resources than that guy. **Panellist:** (L3) Isn't the reality that in the last ten or twenty years there's been a massive transfer of wealth from the poor to the rich, and from the young to the old, and that what you're really trying to do is to justify that by blaming the poor for their position? **Witness:** (L4) No, I'm not trying to justify anything.
DPCR	**Witness:** (L1) I think that all humans should be vegan. (L2) In the sense that a world of veganism would be a more ethical world: its morals would bring benefits to human society. **Panellist:** (L3) Don't you think that killing animals for food is a survival instinct, and so not inherently unethical or morally blameworthy? **Witness:** (L4) I don't think so, I think that instinctive, natural behavior is counterproductive can create problems, both for the individual and society, and both might want it removed. If both deem it immoral and unethical, then it is it as such and the unwanted behaviors should be shied away from and hopefully removed if possible.

Table 18: Examples of the original dialogues MORAL MAZE$_{original}$ and context-enhanced original dialogues MORAL MAZE$^{+context}_{original}$, as well as a dialogue generated with the full DPCR algorithm.

Algorithm	Example of a generated dialogue
DPCR$_{-templ}$	**Witness:** (L1) All humans should be vegan. (L2) A world of veganism would be a more ethical world: its morals would bring benefits to human society. **Panellist:** (L3) Killing animals for food is a survival instinct, and so not inherently unethical or morally blameworthy. **Witness:** (L4) Instinctive, natural behavior is counterproductive can create problems, both for the individual and society, and both might want it removed. If both deem it immoral and unethical, then it is it as such and the unwanted behaviors should be shied away from and hopefully removed if possible.
DPCR^{-rel}	**Witness:** (L1) I believe that Veganism is a natural right. (L2) In the sense that humans sit in the greatest position of control on earth, to rule it and shape it as though the highest power in it. Since we are considering that inalienable rights are endowed by natural law, we must be inferring that there is a natural preference for how justice is shaped. Nature, (particularly the expression of life), is most at peace when ruled in fairness, so it follows that natural law should direct humans to be benevolent. Humans are meant to be vegan. **Panellist:** (L3) Don't you think that there is no evidence proving that humans were created by a mindless force of evolution, and there is overwhelming evidence that many have found the mind of the creator can be reasonably discerned? **Witness:** (L4) I don't think that's true. I think that there is an overwhelming consensus in the scientific community to support the claim all life on earth is the result of Evolution.
DPCR$^{-rel}_{-templ}$	**Witness:** (L1) Veganism is a natural right. (L2) How humans are meant to behave is not necessarily defined by what is best for their human health. **Panellist:** (L3) An abnormal health condition can result in a risk to a person's life if they were to live a normal lifestyle. In context of veganism, if a person's digestive system has become unable to sustain life without eating meat, there is an unnatural conflict between the human's right to live versus the animal's right to live, where the vegan cannot ultimately choose to preserve the life of both. **Witness:** (L4) It should be argued that the right to live with good conscience qualifies the right to take one's own life.

Table 19: Examples of generated dialogues for the algorithms **DPCR**$_{-templ}$, **DPCR**$^{-rel}$ and **DPCR**$^{-rel}_{-templ}$. These three algorithms are ablated version of the full DPCR algorithm.

H Results by topic

Algorithm	Med. Coh.	Mean Coh.
Brexit		
DPCR$_{-templ}$	6	5.91
DPCR	6	5.86
DPCR$^{-rel}_{-templ}$	5	4.64***
DPCR^{-rel}	5	4.50***
Veganism		
DPCR$_{-templ}$	7	6.01
DPCR	6	5.82
DPCR$^{-rel}_{-templ}$	5	4.64***
DPCR^{-rel}	5	4.31***
Vaccination		
DPCR$_{-templ}$	7	6.06
DPCR	6	5.95
DPCR$^{-rel}_{-templ}$	5	4.71***
DPCR^{-rel}	5	4.57***

Table 20: Median and mean coherence scores by topic. *** indicates highly significant differences with DCPR (Student's t-test, $P \leq 0.001$).