Using Discourse Connectives to Test Genre Bias in Masked Language Models

Heidrun Dorgeloh[1], **Lea Kawaletz**[1], **Simon David Stein**[1],
Regina Stodden[2] and **Stefan Conrad**[3]*
[1] Department of English and American Studies – Faculty of Arts and Humanities
[2] Department of Computational Linguistics – Faculty of Arts and Humanities
[3] Department of Computer Science – Faculty of Mathematics and Natural Sciences
Heinrich Heine University Düsseldorf, Germany
{firstname.secondname}@hhu.de

Abstract

This paper presents evidence for an effect of genre on the use of discourse connectives in argumentation. Drawing from discourse processing research on reasoning-based structures, we use fill-mask computation to measure genre-induced expectations of argument realisation, and beta regression to model the probabilities of these realisations against a set of predictors. Contrasting fill-mask probabilities for the presence or absence of a discourse connective in baseline and finetuned language models reveals that genre introduces biases for the realisation of argument structure. These outcomes suggest that cross-domain discourse processing, but also argument mining, should take into account generalisations about specific features, such as connectives, and their probability related to the genre context.[1]

1 Introduction

Argumentative structures in discourse, which comprise a claim and a supporting or attacking premise, exhibit significant variation in their realisation. Notably, as argumentative coherence can be achieved by alternative signals (Cabrio et al., 2013; Das and Taboada, 2018), arguments vary in whether the claim and premise are linked by a discourse connective or not. For example, *because* in item 1a explicitly conveys a causal relation. In contrast, item 1b, where two sentences are separated by punctuation, leaves the relation implicit, with the claim indicated only by the deontic modal *should*.[2]

(1) a. Masking should be mandated *because* it keeps everyone safe.
b. Masking should be mandated. It keeps everyone safe.

The explicit or more implicit realisation of argumentation is a challenge for an understanding of argumentative discourse. However, the processing of arguments is likely not random and should conform with general discourse processing principles. In particular, it can be assumed that, following the Uniform Information Density (UID) hypothesis (Frank and Jaeger, 2008), relations within discourse, such as the one between a claim and a premise, are more likely to be expressed explicitly when they are unexpected, and more likely to be implicit when a relation can be anticipated (Torabi Asr and Demberg, 2012).

The factors that shape expectations in discourse are diverse. Local cues within a phrase or sentence, such as the use of the connective *because* in item 1a, play a role. However, more global forces, such as the overall nature of the document, also drive expectations and, with that, information density (Meister et al., 2021). Knowing that genres guide expectations and influence human discourse understanding on many levels of a text (Giltrow, 2010), we explore in this paper how genre creates a bias for the ways argument structures are realised. These structures are based on relations of subjective causality, a coherence relation that is particularly likely to be driven by contextual signals, including the genre (e.g. Canestrelli et al., 2016; Scholman et al., 2020). For example, for a reader of a newspaper editorial these argumentative structures will be much more expected than for one of a novel or monograph.

Our study compares the predicted presence or absence of discourse connectives in arguments taken from New York Times (NYT) editorials. Due to the UID principle we hypothesise that,

*Contributions: Linguistic background, conception of work and discussion: Heidrun Dorgeloh; conception of work, design and preparation of study: Lea Kawaletz; computational modelling: Regina Stodden; statistical modelling and statistical analysis: Simon David Stein; computational background and discussion: Stefan Conrad. All authors contributed to manuscript revision, read and approved the submitted version.

[1]The data and code for the present study can be found at https://osf.io/n6hq5/.

[2]These are constructed examples based on item 2.

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 27–38

in genres with predictable argumentative structures, such as editorials, there is a lower likelihood of making a relation explicit with a connective. We also assume that an LM finetuned with data from such genres is likely to show a stronger effect of this tendency. To test our hypothesis, we compare baseline (non-finetuned) masked language models (MLMs) with the corresponding finetuned models genre-adapted to editorials. The comparison of models enables us to disregard frequency effects. In this way, the approach allows to verify genre-induced expectations for argument realisation and produces insights which could in the future improve cross-domain discourse processing.

2 Background

2.1 Defining *arguments*

Our understanding of what constitutes argumentative discourse follows established terminology, especially from the field of argument mining (e.g. Stab and Gurevych, 2017; Stede and Schneider, 2018), where an argument, such as exemplified in item 2, consists of two kinds of argumentative discourse units (ADUs): a controversial statement, the *claim* (marked in bold), and another statement which supports or attacks the claim, the *premise* (underlined).

(2) **[M]asking should be mandated and enforced.** It's not just about your individual risk tolerance, but about keeping everyone safe.

ADUs can occur in a single sentence or span multiple sentences, as in item 2. Also, multiple premises may refer to the same claim, forming a single argument. For simplicity, the data analysed for this project only included arguments consisting of one claim and one premise.

2.2 Connectives and discourse relations

Discourse connectives cover the syntactic classes of coordinators (e.g., *and*, *but*), subordinators (e.g., *because*, *while*), as well as connective adjuncts (e.g., *therefore*, *however*) (Dorgeloh and Wanner, 2022). They make the coherence relation between two (or more) ADUs explicit, which is why they are a prominent feature both for studies of discourse coherence and of argumentation structure (Marcu and Echihabi, 2002; Xu et al., 2012; Goudas et al., 2014; Shi and Demberg, 2019; Crible and Demberg, 2020; Kurfalı and Östling, 2021). However, the extent of the actual presence of connectives is often surprisingly low. For example, in the RST Signalling Corpus (Carlson et al., 2002; Das et al., 2015) or the Penn Discourse Treebank (Prasad et al., 2008) – both based on Wall Street Journal texts that in all likelihood contain argumentative texts – more than half of the discourse relations are not marked by a discourse connective. One possible reason for their absence is that there are numerous other options of signalling a coherence relation (Cabrio et al., 2013; Das and Taboada, 2018).

Another reason is that the support or attack relation within arguments has a subjective "source of coherence", that is, the relation does not exist at the propositional content level but at the level of reasoning (Sanders et al., 2021), as in item 2. For these relations, connectives serve as processing instructions, enabling a reader or listener to evaluate how a premise supports or attacks a given claim (Wei et al., 2021a). Psycholinguistic evidence has shown that overly explicit marking of subjective coherence relations triggers a "forewarning effect", alerting the reader to a persuasion attempt (Kamalski et al., 2008). In that sense, connectives can potentially induce resistance against argumentation. Given this effect, it is plausible to assume that argumentative structures are not made more explicit than necessary.

How the needs for explicitness are balanced likely aligns with the UID hypothesis (Frank and Jaeger, 2008). It suggests that discourse relations, including support or attack within arguments, "should be expressed explicitly with a discourse connector when they are unexpected, but may be implicit when the discourse relation can be anticipated" (Torabi Asr and Demberg, 2012, 2669). If expectations are crucial in that sense, a major factor driving explicitness must be the genre, as genres can be seen as schemata "referring to a set of expectations" (Piata, 2016, 255). It follows that, in argumentative texts, such as editorials, the relation between two ADUs is less likely to be expressed with a connective, since the presence of argumentation in this genre can be expected. For illustration, consider item 2 again, where the ADUs are not linked by means of a connective. By contrast, in the adapted variant in item 2′, the argument relation is made explicit by adding the connective *because*.

(2′) **[M]asking should be mandated and enforced** [because] [i]t's not just about your individual risk tolerance, but about keeping everyone safe.

Following the UID hypothesis, item 2′ is the less likely argument pattern in argumentative texts compared to item 2.

2.3 Connectives and language modeling

Argument realisation is a classic issue for the automatic retrieval of arguments, i.e., in argument mining. Connectives, in this context also commonly referred to as *discourse markers*, are seen as indicators of argumentative structure (e.g., Eckle-Kohler et al., 2015; Stab and Gurevych, 2017; Sileo et al., 2019), but "missing" discourse markers are also known to be the rule rather than the exception (Moens, 2018). One reason is that explicitness in argumentation goes beyond using connectives; it also involves other stance markers, as every argument expresses a stance toward its topic (Stein and Wachsmuth, 2019). Connectives and other markers thus together play a role in facilitating the processing of subjective coherence relations (Wei et al., 2021b), but how they interact is still not fully explored. Stodden et al. (2023) also argue that connectives can play a prominent role in stance detection. They extract the probabilities of connectives for a claim-premise relation from a MLM and show that training a simple classifier using these values as features is capable of optimising stance detection. Our approach here uses a similar line of research.

Another reason why the presence or absence of discourse connectives as indicators of arguments is not fully understood is the lack of cross-genre generalisations. In a recent paper, Rocha et al. (2023) report that introducing connectives as signals of the relation between a claim and a premise has the potential to improve argument mining. They employ finetuned LMs trained on both real and constructed arguments to introduce connectives between ADUs, which improves cross-genre transfer. However, their approach does not consider genre-specific associations of explicit indicators like connectives and the context. To address this, we aim to incorporate genre generalisations through genre-induced fine-tuning.

Our approach is to explore the presence or absence of a discourse connective for the claim-premise relation in arguments using BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). In the task of these MLMs, the objective is akin to a cloze test; the model learns to predict words for randomly masked tokens in the original input texts (Devlin et al., 2019). Unlike a causal LM, which predicts the next word solely based on the previous context, an MLM can predict a word in the middle of a sequence based on both left and right context. We use the cloze position between the claim and premise by extracting probabilities for connectives and, as a proxy for their absence, punctuation marks. In doing so, we refrain from using causal LM prompting methods and instead compare the probabilities of different types of marking in a statistical analysis, which necessitates more than just listing the top n markers.

2.4 Hypotheses

We compare the predictions made by different models, first for the difference between explicit connective and no marking and, then, for the comparison between baseline models and models finetuned for editorials. In this context, we make three predictions. First, given that newspaper editorials are a genre whose primary goal it is to persuade and which are therefore "one of the purest forms of argumentative text" (Al-Khatib et al., 2016, 3440), the discourse relations that are characteristic of this type of discourse are subjective causal relations, i.e., discourse relations that do not refer to propositional content, but to reasoning (see subsection 2.2). Due to the forewarning effect, we assume that these relations are not marked more explicitly than necessary (→ H1 below). Second, if a genre as a whole involves argumentation, the UID hypothesis suggests that argument relations are expressed more implicitly in this genre than in other, less persuasive genres. It follows that, in LMs finetuned on strongly persuasive discourse, argumentation is even more likely to occur without a connective (→ H2). Third, regarding the magnitude of this effect, we predict that it depends on the models' baselines, given their varied training data. LMs trained on comparatively non-argumentative texts (e.g., books and Wikipedia, for BERT) should show a more pronounced difference between finetuned and baseline versions than those whose training already included a certain proportion of texts from more argumentative genres (e.g., news and web-based texts, for RoBERTa; → H3).

H1 The absence of a connective (here: indicated by a punctuation marker) is more likely than the presence of an explicit discourse connective.

H2 LMs that have been finetuned on argumentative genres (here: editorials) predict a lower probability for a discourse connective than the baseline ones.

H3 This effect is more pronounced in LMs trained on non-argumentative texts (here: BERT) than in those trained on a larger portion of argumentative texts (here: RoBERTa).

3 Methodology

Our method involves comparing explicit to non-explicit realisations of the claim-premise relation in a set of arguments. We use different LMs to quantify the acceptability of the presence of a discourse connective as probabilities of the masked-tokens. These can be seen as a placeholder for the realisations in MLMs.

The bidirectional architecture of these models enables the prediction of token probabilities based on both ADUs (claim and premise). This prediction is dependent on the training data of an LM. To adapt the model to a genre of argumentative texts, we finetuned the LM on additional NYT editorials which are not part of our annotated data, which enables us to compare finetuned with non-finetuned models.

3.1 Data

The data set was manually selected with the aim to test this new approach for exploring genre generalisations. The data set consists of 81 arguments from a corpus of 2,508 NYT editorials (3,227,122 tokens). These were published between January 2020 and June 2021 with at least one of the NYT tags 'coronavirus (2019-ncov),' 'vaccination and immunization,' or 'epidemics.' The selection followed a "purposeful sampling" approach (Patton, 2015), which means we did not aim for a representative sample of all arguments attested in the corpus. Instead, we identified arguments in a subset of 50 editorials (55,603 tokens) and chose 81 arguments in an elaborate and resource-intensive process tailored towards the proof-of-concept nature of our analysis. The process took place in several steps that we describe in detail in our guidelines for annotation (Kawaletz et al., 2023). The selection of arguments was based on the following principles: Not only did all arguments have to adhere to the semantic classification of arguments we have developed (Kawaletz et al., 2022), but they were, at a minimum, identified by two out of three annotators, and subsequently confirmed by two curators, all possessing linguistic training.

Table 1 provides a summary of the data set properties, outlining the features that were integrated into our statistical analysis (Kawaletz et al., 2022): connective (are claim and premise connected by a connective?), relation (does the premise support or attack the claim?), and category (does the claim state that something is or is not the case, or does it mandate an action or prohibition, or does it evaluate something positively or negatively?). As expected, most claim-premise pairs lack a connective (74.07%), reflecting the tendency of argumentative discourse to favour implicit relations (see subsection 2.2). It also becomes obvious that support relations dominate (86.42%), and that most arguments in the data set are epistemic in nature (71.60%).

Property	Option	Count	Per cent
Connective	Present	60	74.07%
	Absent	21	25.93%
Relation	Support	70	86.42%
	Attack	11	13.58%
Category	Epistemic	58	71.60%
	Deontic	19	23.46%
	Ethical	4	4.94%

Table 1: Properties of the data set

Finally, the arguments span a broad range of lengths, from the shortest at 11 words to the longest at 90 words, with an average of approximately 44.05 words and a median of 42 words.

3.2 Extraction of probabilities

In order to calculate the probability for the presence or the absence of a connective, we conducted the following preprocessing steps: i) The last character from ADU1 is truncated to prevent the punctuation character from affecting the predictions. ii) If ADU2 starts with a connective , the connective of ADU2 is truncated to prevent the concatenation of two connectives in a row or of a connective and punctuation mark.

Connectives		Punctuation markers	
although	unless	**.**	-
because	while	**;**	–
but	yet	**,**	—
since	anyway^{B}	**:**	...
so	consequently^{B}	**?**	. . .
still	hence^{B}	**—**	!
and	however^{B}		–^{B}
as	nevertheless^{B}		..^{R}
for	therefore^{B}		
thus	whereas^{B}		

Table 2: Presence or absence of explicit marker queried in the LMs' output. Bold face markers also occur in the NYT data set. Markers with ^{B} are used only for BERT, while those with ^{R} are exclusive to RoBERTa.

Next, both ADUs were concatenated with model-specific masked tokens: *[MASK]* for BERT and <*mask*> for RoBERTa. For instance, item 2 was input to BERT as in item 3.

(3) Masking should be mandated and enforced [MASK] it's not just about your individual risk tolerance, but about keeping everyone safe.

We calculated the probabilities of the masked-token for each possible token (or subword) with a Python pipeline for "fill-mask" included in the Huggingface `transformers` package (Wolf et al., 2020).[3] As opposed to the approach of Rocha et al. (2023), the method does not involve filling the gap between claim and premise with an explicit marker, but at extracting the probabilities of a list of tokens.

From the resulting probabilities list, we extracted the probabilities of 34 tokens of interest (see Table 2)—20 discourse markers (for explicit realisations) and 14 punctuation marks (indicating the absence of a connective). A connective was added to the list of explicit markers if it is a single-word connective, and i) a coordinating or subordinating conjunction expressing a support or attack relation, or ii) a "linking adverbial" (Biber et al., 2021, 755) expressing a support or attack relation. A punctuation mark was added to the list if it occurs in our masked data, and/or if it was in the list of the top 10 predicted tokens of the LM using our data.

We did not include multi-token connectives (e.g., *for this reason* or *on the other hand*) as the fill-mask approach is only available for one-(sub)token prediction. Compound connectives had to be excluded because most LMs are using subword tokenizers, hence, they would be split into several subtokens (e.g., *anyway* would be tokenized as *any* and *way*) and cannot be predicted as a whole token in the fill-mask task.[4]

3.3 Language models and finetuning

For our experiments, we chose BERT-large-uncased (Devlin et al., 2019) and a derivative model, RoBERTa-large (Liu et al., 2019).[5] While sharing the same architecture they are pre-trained on different genres: BERT is pre-trained on 16 GB of data from English books and Wikipedia, whereas RoBERTa is pre-trained additionally on 144 GB of news and web texts. Our selection of these specific LMs was driven by a focus on the impact of genre. However, the differences between BERT and RoBERTa extend beyond their training data. For instance, a) RoBERTa is solely trained for language modelling, unlike BERT, which also includes next sentence prediction; b) they employ different tokenisation methods: RoBERTa uses Byte-Pair Encoding, while BERT uses WordPiece; c) RoBERTa is case-sensitive, whereas the version of BERT we chose is not. Despite these variations, BERT and RoBERTa were the most suitable models for our research objectives. For example, XLM-RoBERTa-base (Conneau et al., 2020) shares the same architecture as BERT and RoBERTa, but includes multilingual training, and DistilBERT-base-uncased (Sanh et al., 2019) is trained on the same data as BERT, but has fewer tunable parameters.

We then applied *domain-adaptive finetuning* (Han and Eisenstein, 2019), an unsupervised method that adapts the LM to a new or underrepresented genre. We chose this approach to adapt the LMs for argumentative texts because they are primarily trained on non-argumentative data while also incorporating argumentative data to varying extents. Specifically, we fine-

[3] The determination of the probabilities is limited to the top_k, where k is the length of the vocabulary. Following this, some probabilities are close to 0 (very unlikely).

[4] We are comparing models with the same tokenizer, i.e., the baseline model and the finetuned model. Hence, for a different set of connectives, we would not expect a strong effect on our results.

[5] As previously mentioned, we are not using autoregressive LMs (e.g., ChatGPT or Llama) and prompting methods as we are interested in the probabilites of different types of marking for further statistical analysis. MLMs have the advantage over autoregressive LMs to provide the probabilities of a word at any position within a sequence by considering both the left and right context, rather than solely predicting next words at the end of a sequence.

tuned on the 2,458 NYT editorials from our corpus (but excluding those 50 from which we selected the arguments for our data set). This way, the finetuned LMs are more likely to mirror the lower likelihood of a connective for editorials and, in that, for an argumentative genre.[6]

3.4 Statistical analysis

We fitted generalised additive models of the beta regression family to the data, using the `mcgv` package (Wood, 2017) in `R` (R Core Team, 2023). Beta regression is uniquely suited to model proportional values (see, e.g., Ferrari and Cribari-Neto, 2004). These models also allow us to include a number of important control variables.

Response variable We name our response variable PROBABILITY, referring to the probability of masked tokens estimated by the LMs. For each argument in our data set we calculated two probability measurements, one for the presence of an explicit discourse marker and one for its absence. The probability of an explicit discourse marker was calculated by taking the sum of the estimated probabilities of all connectives. The probability of the absence of marking was the sum of the estimated probabilities of all punctuation marks. Each of these two measurements was paired with the value `present` or `absent` in an additional variable CONNECTIVE. This coding enables us to investigate both types of probabilities in a single statistical model.

Predictor variables Our two predictor variables of interest are CONNECTIVE and MODEL. CONNECTIVE specifies whether we look at the probability for the presence or absence of explicit marking. MODEL specifies which LM estimated these probabilities: `baseline BERT`, `finetuned BERT`, `baseline RoBERTa`, or `finetuned RoBERTa`.

We use the control variable N_TOKENS, the number of word tokens in the sentence, to gauge sentence length and complexity. It may be expected that longer and more complex sentences will exert greater pressure to use punctuation marks, thereby disfavouring marking.

Additionally, we control for RELATION and CATEGORY. RELATION specifies the relation between premise and claim (`attack` or `support`). We expect that, compared to support relations, attack relations favour explicit marking, since contrasting relations are cognitively more complex, requiring more cues (Crible and Demberg, 2020). CATEGORY specifies the semantic argument category (`epistemic`, `ethical`, or `deontic`). We expect deontic arguments to exhibit the strongest dispreference for explicit marking because claims demanding an action often contain a deontic modal, expressing necessity (e.g., *should*), which already implies the presence of a premise (Kawaletz et al., 2023).

Furthermore, we specify with HASNECESSITYMODAL and HASDEMDET whether the sentence contains at least one necessity modal (*must*, *should*, or *ought*) or at least one demonstrative determiner (e.g., *this*, *these*), respectively. Both are features that could reduce the likelihood of explicitness by way of a connective, as they are also known to be linguistic features of persuasion and argumentation (Biber, 1989; Petch-Tyson, 2000).

Finally, we include SOURCEID, the identifier of the source document of the target sentence, to control for potential variation in probabilities introduced by different authors or texts.

Modelling We fitted six types of beta regression model: i) one for baseline BERT, ii) one for finetuned BERT, iii) one for baseline RoBERTa, iv) one for finetuned RoBERTa, v) one that compares baseline BERT and finetuned BERT, and vi) one that compares baseline RoBERTa and finetuned RoBERTa. The first four types of model investigate the difference in probability between the presence or absence of explicit marking for each LM individually. They do not include MODEL as a predictor. Models v and vi investigate the difference between finetuned and baseline models. They include MODEL as a predictor of interest.

We fitted each type of model as a simple version and a complex version. The simple versions include only the predictors of interest (CONNECTIVE for the four individual models and an interaction of CONNECTIVE and MODEL for the two comparisons). The complex versions include interactions of CONNECTIVE with each of the covariates described above (SOURCEID was not included in an interaction).

Following standard procedure, we reduced the models by removing non-significant terms (at the .05 alpha level) in a stepwise fashion

[6] You can find the hyperparameters in Appendix A and the code with more details in the osf repository.

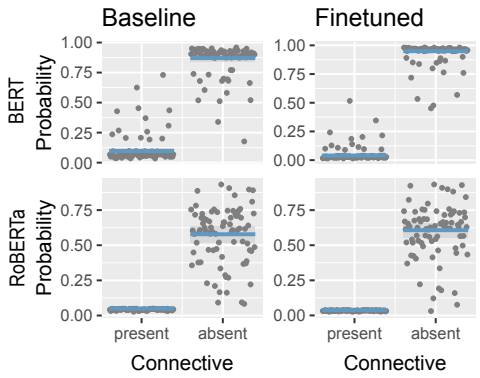

Figure 1: Probability of the presence and absence of explicit argument marking for the four LMs.

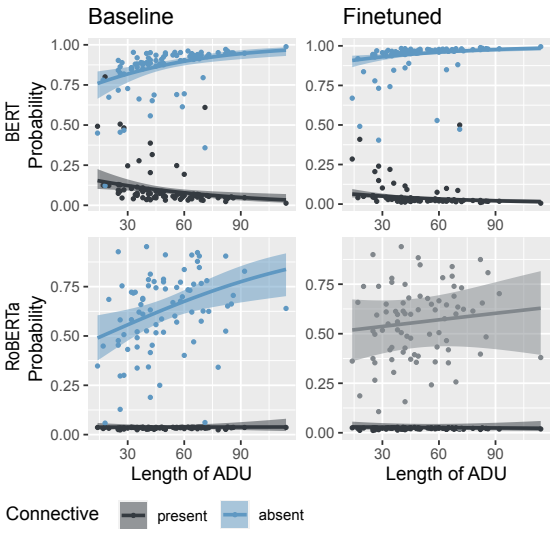

Figure 2: Interaction of CONNECTIVE with sentence length for the four LMs. Greyed out effects did not reach significance and were eliminated in the statistical model.

(highest p-value first) until only predictors remained of which at least one level reached significance.

4 Results

Figure 1 plots the results of the simple versions of the four models, which predict the probabilities of explicit marking being present or absent for each LM individually.[7] This reality check confirms our expectation that explicit marking is disfavoured across models. In all four cases, we find very highly significant effects (at $p < .001$) of CONNECTIVE on PROBABILITY in the expected direction. Note that this effect is likely in part a frequency artefact. The proxy measure by which we gauge the absence of marking, i.e., punctuation, will naturally yield higher probabilities than the proxy by which we measure explicit marking , i.e., connectives.

In the complex versions of these four individual models, which include interactions of the predictors with CONNECTIVE, the interactions and main effects of the covariates mostly do not reach significance. One exception is N_TOKENS. Figure 2 shows that in three out of four complex models our expectations are confirmed: With increasing ADU length, the absence of marking becomes even more probable, while explicit marking becomes even less probable. In some models we also find the occasional expected interaction with other covariates, such as RELATION: Support relations feature even higher probabilities for `absent` or even lower probabilities for `present` compared to attack relations. Details can be found in the supplementary materials.

[7]The interested reader can view all full models in the supplementary materials at the osf repository.

Let us now turn to the question of genre, i.e., the comparison of baseline LMs with finetuned LMs. Figure 3 plots the main result from each of the two complex beta regression models, the top panel showing the interaction of CONNECTIVE with the BERT LMs, the bottom panel showing its interaction with the RoBERTa LMs (the results are the same in the two simple versions of each regression model).

Both BERT LMs disfavour explicit marking, but the finetuned version prefers such marking to be absent significantly more than does the baseline version of BERT. Again, frequency effects likely amplify the strong dispreference for connectives. However, a general bias for punctuation exists for both baseline and finetuned LMs, enabling us to compare them directly. Moving down to the bottom panel, we can observe that RoBERTa, too, prefers the absence of explicit marking even more when finetuned, but here, the effect fails to reach significance. As it is difficult to interpret the absence of an effect in the frequentist framework, we used the BIC approximation to the Bayes Factor (Wagenmakers, 2007) to compare the model for RoBERTa against a null hypothesis model without MODEL and its interaction with CONNECTIVE. This analysis indicates that the data are more likely under the null hypothesis (finetuning RoBERTa does not affect the presence of a connective) than under the hypothesis (finetuning RoBERTa does affect the presence of a connective) ($BF_{01} = 32.79$). If we assume that it is

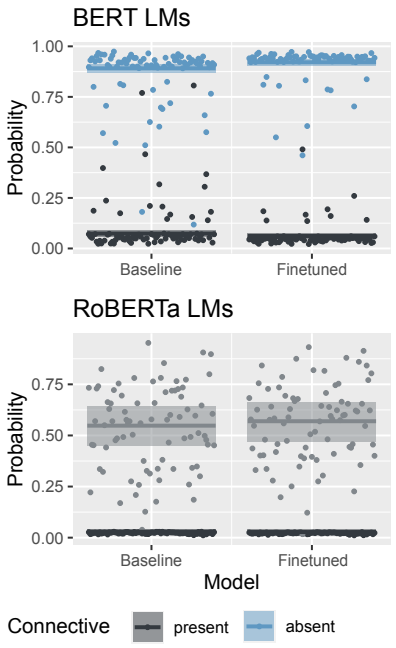

Figure 3: Interactions of CONNECTIVE with MODEL, each comparing the baseline version of the model with the finetuned version. Greyed out effects did not reach significance and were eliminated in the statistical model.

a priori equally likely that finetuning RoBERTa does and does not have an effect, the posterior probability we find ($Pr_{H_0}|D = .97$) constitutes "strong" evidence for the null, according to the Raftery (1995) classification scheme. We can thus be confident that while we find a finetuning effect for BERT, we are dealing with a true null result for RoBERTa.

5 Discussion

The LMs have shown a clear preference for the absence of a discourse connective, which overall confirms a characteristic of argument structures, i.e. their subjective causal relations, in line with the psycholinguistic background to our approach (H1). Also, in line with our expectations, after finetuning, the LMs both showed a decreased probability for an explicit connective compared to the baseline ones (H2).

However, only BERT, but not RoBERTa, shows a clear, i.e., statistically significant, increase. We believe the fact that we find a significant finetuning effect for BERT but not for RoBERTa is a true effect of genre (H3): The baseline version of BERT was trained on less argumentative texts (specifically, books and Wikipedia only) compared to RoBERTa, which also includes news and websites in its training data. The increase in the "argumentativeness" of genres from baseline to finetuned is thus higher for BERT than for RoBERTa, which has already seen many argumentative texts before having been finetuned. For RoBERTa, then, the finetuning effect is less pronounced. This difference that we observed is in line with the assumption that genre does create a bias for the realisation of argument structure in discourse.

Several effects we have presented suggest that the approach covers the use of connectives and genre conditions, as far as they are identifiable for an LM, reasonably well. The fact that the absence of a connective is highly likely across all models (H1) is likely a frequency effect. However, we were able to disregard this effect by focusing on the comparison of baseline and finetuned versions (H3), since it applies to both equally. Our modelling also showed that, in line with expectations, absence of a connective becomes overall more likely with increased sentence length (number of tokens) – a finding which suggests that length is not only a control variable, in the sense of reflecting the complexity of the pairing of claim and premise. The effect of length also suggests that there are other features relevant for the explicitness of an argument, and their presence will become more likely the longer an argument gets. For example, other markers known to typically connect an argument's second constituent are features at the sentence beginning, the so-called "theme zone", such as adjuncts or demonstrative expressions (Fetzer, 2018; Petch-Tyson, 2000). In general, a clear effect of overall length of the discourse units confirms the relevance of information density for the use of connectives.

Our results also indicate that argument mining could profit from genre generalisations. So far, approaches are typically developed and trained using data either from one genre (e.g., persuasive essays in Stab and Gurevych, 2017) or mixed-genre corpora with no systematic cross-genre transfer (e.g., Morio et al., 2022). In the former case, while high accuracy is often achieved within the same genre, the transfer to another genre usually weakens the results. In the latter case, the approach often achieves moderate accuracy across genres without excelling in any specific genre. This indicates a general oversight of genre as a systematic factor in current methodologies. However, genre generalisations are crucial for dealing with a potential problem of LMs when dealing with ar-

gumentative discourse: the genre-specific use of explicit marking may lead LMs to learn only the markings used within the genre(s) available in the training data, thereby possibly overlooking or neglecting other patterns.

6 Conclusion and outlook

In this study, we have presented a method for testing genre bias in LMs, and we have shown that discourse expectations as driven by the genre have an impact on the explicit linking of ADUs by way of discourse connectives. We used two discourse processing principles – forewarning and UID – to account for a general preference for the absence of a connective. Testing this preference in the form of fill-mask probabilities of our LMs enabled us to identify an expected genre bias after finetuning.

Even if the computational approach piloted with this work is not without its limitations – being based on a very small data set and focusing solely on single-word connectives while excluding other discourse markers – it successfully quantifies the influence of genre on discourse-structure realisation. In that sense, the method can serve as a role model for investigating genre effects. However, most argumentative discourse will contain many other cues for realising argumentation, which aligns with the identified effect of argument length. Extending the approach to multi-word connectives, to combinations of connectives and punctuation, or to more complex "alternative lexicalizations" that equally express coherence relations (Knaebel and Stede, 2022) would therefore be a promising endeavour. In addition, from a computational standpoint, it would be beneficial to apply our approach to other LMs, particularly considering that only BERT, not RoBERTa, incorporates next-sentence prediction.

Overall, this work shows that both cross-domain discourse processing and argument mining can benefit from genre generalisations. While recent work in argument mining has aimed at making LMs less genre-dependent by way of using connectives (Rocha et al., 2023), our approach highlights a method of revealing genre bias in the use of connectives and could thus be a template for future, more genre-dependent work.

References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Douglas Biber. 1989. A typology of English texts. *Linguistics*, 27(1):3–44.

Douglas Biber, Stig Johansson, Geoffrey N. Leech, Susan Conrad, and Edward Finegan. 2021. *Grammar of Spoken and Written English*. John Benjamins.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *Computational Logic in Multi-Agent Systems*, volume 8143, pages 1–17. Springer, Berlin, Heidelberg.

Anneloes Canestrelli, Pim Mak, and Ted Sanders. 2016. The influence of genre on the processing of objective and subjective causal relations: Evidence from eye-tracking. In Ninke Stukker, Wilbert Spooren, and Gerard Steen, editors, *Genre in Language, Discourse and Cognition*, pages 51–74. De Gruyter.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank LDC2002T07*. Linguistic Data Consortium, Philadelphia. Web Download.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ludivine Crible and Vera Demberg. 2020. The role of non-connective discourse cues and their interaction with connectives. *Pragmatics & Cognition*, 27(2):313–338.

Debopam Das and Maite Taboada. 2018. Signalling of Coherence Relations in Discourse, Beyond Discourse Markers. *Discourse Processes*, 55(8):743–770.

Debopam Das, Maite Taboada, and Paul McFetridge. 2015. *RST Signalling Corpus LDC2015T10*. Linguistic Data Consortium, Philadelphia. Web Download.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Heidrun Dorgeloh and Anja Wanner. 2022. *Discourse Syntax: English Grammar Beyond the Sentence*. Cambridge University Press.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal. Association for Computational Linguistics.

Silvia Ferrari and Francisco Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Anita Fetzer. 2018. The encoding and signalling of discourse relations in argumentative discourse. *The construction of discourse as verbal interaction*, pages 13–44.

Austin F. Frank and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 911–916, Washington, DC, USA. Cognitive Science Society.

Janet Giltrow. 2010. Genre as difference: The sociality of linguistic variation. In Heidrun Dorgeloh and Anja Wanner, editors, *Syntactic Variation and Genre*, volume 70, pages 29–52. DE GRUYTER MOUTON, Berlin, New York. Series Title: Topics in English Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham. Springer International Publishing.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Judith Kamalski, Leo Lentz, Ted Sanders, and Rolf A. Zwaan. 2008. The forewarning effect of coherence markers in persuasive discourse: Evidence from persuasion and processing. *Discourse Processes*, 45(6):545–579.

Lea Kawaletz, Heidrun Dorgeloh, and Stefan Conrad. 2023. Annotation guidelines for the project *Probing patterns of argumentative discourse*. Manuscript.

Lea Kawaletz, Heidrun Dorgeloh, Stefan Conrad, and Zeljko Bekcic. 2022. Developing an argument annotation scheme based on a semantic classification of arguments. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 62–67, Edinburgh, UK. Association for Computational Linguistics.

René Knaebel and Manfred Stede. 2022. Towards identifying alternative-lexicalization signals of discourse relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 837–850, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 1–10, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*, abs/1907.11692. Version 1.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marie-Francine Moens. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*, 9(1):1 – 14.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end argument mining with cross-corpora multi-task learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

Michael Quinn Patton. 2015. *Qualitative research & evaluation methods: integrating theory and practice*, fourth edition edition. SAGE Publications, Inc, Thousand Oaks, California.

Stephanie Petch-Tyson. 2000. Demonstrative expressions in argumentative discourse. *Corpus-Based and Computational Approaches to Dis-*

course Anaphora, Studies in Corpus Linguistics, 3:43–64.

Anna Piata. 2016. *Genre "out of the box": A conceptual integration analysis of poetic discourse*, pages 225–250. De Gruyter Mouton, Berlin, Boston.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Adrian E. Raftery. 1995. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Gil Rocha, Henrique Lopes Cardoso, Jonas Belouadi, and Steffen Eger. 2023. Cross-genre argument mining: Can language models automatically fill in missing discourse markers? *arXiv*, abs/2306.04314. Version 1.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Merel C. J. Scholman, Vera Demberg, and Ted J. M. Sanders. 2020. Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, 57(10):844–861.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Benno Stein and Henning Wachsmuth, editors. 2019. *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy.

Regina Stodden, Laura Kallmeyer, Lea Kawaletz, and Heidrun Dorgeloh. 2023. Using masked language model probabilities of connectives for stance detection in English discourse. In *Proceedings of the 10th Workshop on Argument Mining*, pages 11–18, Singapore. Association for Computational Linguistics.

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India. The COLING 2012 Organizing Committee.

Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804.

Yipu Wei, Jacqueline Evers-Vermeul, Ted M. Sanders, and Willem M. Mak. 2021a. The role of connectives and stance markers in the processing of subjective causal relations. *Discourse Processes*, 58(8):766–786.

Yipu Wei, Jacqueline Evers-Vermeul, Ted M. Sanders, and Willem M. Mak. 2021b. The role of connectives and stance markers in the processing of subjective causal relations. *Discourse Processes*, 58(8):766–786.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Simon N. Wood. 2017. *Generalized additive models*, 2 edition. Chapman and Hall, Boca Raton.

Yu Xu, Man Lan, Yue Lu, Zheng Yu Niu, and Chew Lim Tan. 2012. Connective prediction using machine learning for implicit discourse relation classification. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

A More details on finetuning

hyperparameter	value
mlm probability	0.15
batch size	32
learning rate	0.00002
weight decay	0.01

Table 3: Hyperparameters for finetuning