A practical perspective on connective generation

Frances Yung[1], Merel C.J. Scholman[1], and **Vera Demberg[1,2]**
[1]Language Science and Technology, [2]Computer Science, Saarland University
Saarbrücken, Germany
{frances,m.c.j.scholman,vera}@coli.uni-saarland.de

Abstract

In data-driven natural language generation, we typically know what relation should be expressed and need to select a connective to lexicalize it. In the current contribution, we analyse whether a sophisticated connective generation module is necessary to select a connective, or whether this can be solved with simple methods, such as random choice between connectives that are known to express a given relation, or usage of a generic language model. Comparing these methods to the distributions of connective choices from a human connective insertion task, we find mixed results: for some relations, it is acceptable to lexicalize them using any of the connectives that mark this relation. However, for other relations (temporals, concessives) either a more detailed relation distinction needs to be introduced, or a more sophisticated connective choice module would be necessary.

1 Introduction

We assume a natural language generation setting in which we know, based on content planning, what coherence relation we wish to express, and the problem consists of choosing a suitable connective to express this relation. Of course, this problem is not independent of the discourse relation framework that is used – in the current study, we work with the relation distinctions proposed by PDTB-3 (Webber et al., 2019). From a generation perspective, the task would be easiest if, within relation distinctions, the connectives would be substitutable with each other. For instance, consider a framework that distinguishes causal relations from others, but within this class does not distinguish reason relations from result relations. Using such a framework, we would not be able to correctly select between connectives such as "because" and "therefore" as we wouldn't be able to express their difference. In this case, our connective insertion method would have to be able to learn this distinction in order to allow for fitting choices.

In this study, we aim to answer the following questions:

1. Does the discourse relation hierarchy of PDTB-3 make sufficiently fine-grained distinctions, such that choosing any one of the connectives that can express the relation will lead to a coherent text, or are finer-grained distinctions necessary?

2. Is there a practical value in developing sophisticated methods for connective choice, or are comprehenders largely insensitive to the choice among meaning-equivalent connectives?

We conduct two empirical connective insertion studies to test whether humans have a preference for the connective that was originally present in the text when being asked to lexicalize a relation. Several findings seem possible:

a) All connectives that express a specific coherence relation are fully interchangeable – in this case, we expect to see that human participants do not have any specific preference among the connectives that express the relation. We could conclude that the PDTB distinctions are sufficient, and using a simple connective choice method is sufficient.

b) All connectives are correct, but there are other factors, such as information-theoretic aspects, that influence which connective is preferred (e.g., a short vs. a longer / rarer connective). Again, we would conclude that PDTB distinctions are sufficient, but that other factors such as information density need to be taken into account.

c) Some instances of connectives would not be good choices for lexicalizing a specific in-

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 72–83
November 10–11, 2021.

stance of a relation, even though they can express other instances of that relation – in this case, a relevant distinction in relation sense may be missing or additional features may have to be learned in order to choose a fitting connective.

In both cases b) and c), we would expect to find that humans show a peaked distribution, preferring a specific connective or dispreferring a specific connective to express a relation. A more detailed analysis on these cases is conducted in order to check whether the preference is due to a lack of substitutability or rather a softer factor.

Our first study uses as material a naturalistic distribution of relations and finds that random choice among fitting connectives would achieve good accuracy on this problem – however, the coherence relation distribution in this first study is dominated by a small number of frequent relations. The second study therefore uses a more balanced design to better represent less frequent coherence relations. In this study, we find more nuanced results: while there are indeed some relations for which any of the matching connectives can be inserted, and a language model like GPT-2 (Radford et al., 2019) would perform well, there are also some relations for which simple automatic methods would systematically choose unsuitable lexicalizations. We analyse the latter cases in more detail in section 5.2.

2 Related work

There has been continuing interest in the task of connective prediction in recent years, but mostly as an auxiliary task to coherence relation classification. Zhou et al. (2010) use a language model for the task of connective prediction and Xu et al. (2012) deploy word pairs as well as a set of linguistically motivated features. More recently, Qin et al. (2017); Shi and Demberg (2019) and Kurfalı and Östling (2021) have used connective prediction for implicit relations as a secondary or adversarial task to improve discourse relation classification. The current study tackles a different task than these studies, where the relation label is assumed not to be available. Instead, we assume that the generation system knows what relation should be conveyed, and the remaining problem is the lexicalization using a connective.

Another related study is Ko and Li (2020), who analyse the discourse abilities of GPT-2 and find that connectives are sometimes incorrect. They propose to use a specific discourse component to address connective generation. Again, however, their setting is different from ours, as we assume that connectives need to be inserted into a text which is generated with a known discourse intention.

An experimental study closely related to the current contribution was conducted by Malmi et al. (2018). Crowd-workers were asked to guess the original connective in a text where the explicit connective had been removed. Our analysis of their results shows that the omission of an explicit connective often leads to a change in interpretation of the relation: for ca. 80% of explicitly marked relations, participants did not recover the original connective. Therefore, our study is designed differently: we ensure that workers express the intended relation by providing them with a choice among connectives that can mark the original relation.

The substitutability of connectives has been studied in previous literature. Most notably, Knott (1996); Knott et al. (2002) explored this topic using a connective substitution task, and created a hierarchy of connectives based on these results. Our study adds quantitative data on connective insertion preferences, as well as a practical perspective by investigating connective choice based on PDTB-3 discourse relation labels (and not just substitutability with respect to other connectives).

3 Methodology

We use crowdsourced human experiments to examine whether a specific DC is preferred for a given discourse relation instance. The approach is straight-forward: from discourse-annotated corpora, we sample a set of explicitly marked discourse relations. The original connectives are then removed from the instances and crowdworkers are asked to select the best-suiting connective. Several options of connectives, which are all valid explicit markers of the annotated discourse relation sense, are given to the workers to choose from.

If the choices of connectives made by the crowdworkers reproduce the original connective in the data, it indicates that this particular connective should be preferred over the other alternatives for that specific discourse relation instance. On the other hand, if the choices made by the crowdworkers are evenly distributed per instance, it indicates that the discourse relation instance in question could be interchangeably marked by alternative

connectives.

In addition, we test whether a language model (here, GPT-2, Radford et al., 2019) can select the appropriate connective from the same set of options also given to the human participants. The preferred connective is chosen based on the cross entropy loss of the language model. If the connective with the highest probability based on the language model is identical to the preferred choice by humans, this suggests that a language model is sufficient for generating an appropriate connective for a given discourse relation. If, on the other hand, the language model prefers a connective which is not preferred by humans, this indicates that the language model is missing a relevant aspect of the coherence relation.

3.1 Data

We used discourse-annotated data where each explicit connective is labelled with a discourse relation type. Two datasets were used: 1) the complete English part of the TED Multilingual Discourse Bank (TED-MDB) (Zeyrek et al., 2019); 2) a balanced sample of the Penn Discourse Treebank 3.0 (PDTB-3; Prasad et al., 2018).

The English portion of the TED-MDB consists of the transcription of six English TED talks, which are videos of presentations on various topics. The transcription is annotated with discourse structure following the annotation scheme of PDTB-3 (Webber et al., 2019). Our focus are the explicit connectives and their annotated discourse relation types.

Since our objective is to evaluate the preference for alternative connectives, we exclude cases where the acceptability of the connective is highly restrictive. The majority of such cases is where two verb phrases are linked by a coordinating conjunction. For example, it is not grammatically acceptable to mark the CONJUNCTION relation with other connectives, e.g. *in addition* or *also*, instead of the original *and* in the sentence *"We have a population that is both growing* ***and*** *aging"*. Other items that were removed from our experiment include pragmatic markers (e.g. ***but*** *let's move yet again...*), prepositions (e.g. ***for*** *committing these so-called crimes*), and annotation errors (e.g. ***so*** *awesome*).

From the original set of 290 annotated explicit connectives, 210 are included in the experiment after screening. The distribution of the relation and connective types are shown in Table 3 in the appendix. The options of alternative connectives given to the crowdworkers primarily include the connectives used for the same relation type in the same dataset. Additional common connectives are included such that there are three to five options for each question.

The items from TED-MDB were divided into 4 batches, each consisting of 1 to 2 talks and approximately 50 items. They were presented to the crowdworkers paragraph by paragraph in the same order as the original data (see Figure 1).

The distributions of the relations and connective types in the TED-MDB are highly skewed. While this is representative of distributions found in natural language, it also means that we have very few observations for some relation types. We thus ran an additional study using the same setup on a balanced sample, in order to assess the preference of connectives for less frequent relations.

The balanced sample from PDTB-3 consists of the 12 most frequent explicit relation types in PDTB-3. For each of these, we selected approximately 18 items per relation, resulting in a set of 206 items. We furthermore balanced the number of items marked by different connectives for each relation type. For instance, the ARG2-AS-DENIER relation can be marked using a range of different connectives, including *but*, *however*, *though*, *still* and *yet* as the most frequent ones. We selected the most common connectives with respect to the relation type and included a similar number of instances of relations marked with each of these five connectives in our study. The instances were selected randomly, except for instances with highly restrictive connective usage (as discussed above), instances with multiple labelled relation types and instances where the connective is embedded in the middle of the argument span (because this restricts substitutability). The distribution of the relations and connectives is shown in Table 4 in the appendix.

In the human connective choice experiment, the top 5 most frequent explicit connectives for each relation type are used as connective options. The items from PDTB-3 were presented to workers including one or two sentences of context before and after the discourse relational arguments, depending on sentence length. The items were randomly divided into 8 batches of about 28 items each. Smaller batch sizes were used in this experiment because the workers needed to read more context per item than for the TED-MDB (since the PDTB-3 items do not represent consecutive text).

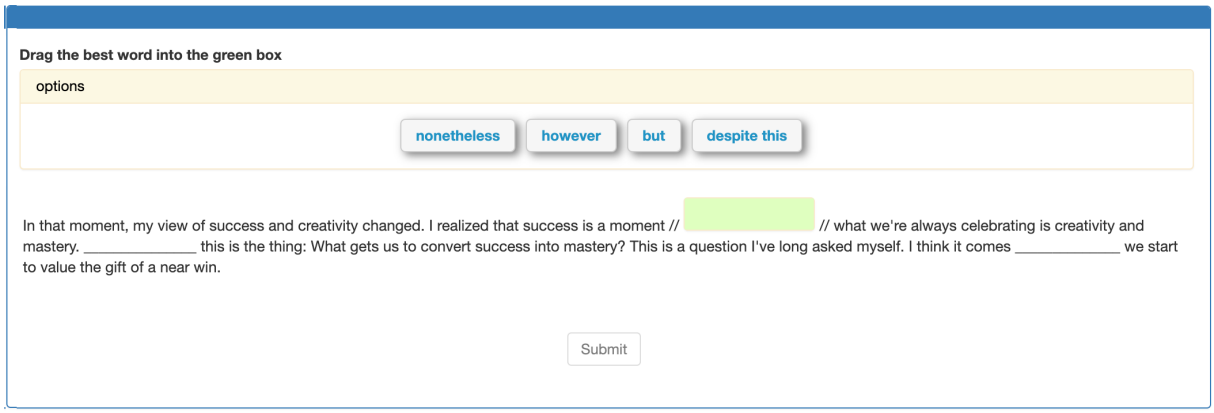

Figure 1: A screenshot of the human evaluation task.

3.2 Procedure

The connective selection task was implemented as a drag-and-drop task, as shown in Figure 1, on the LingoTurk platform (Pusse et al., 2016) and hosted on Prolific. The order of the connective options was randomized. We obtained 10 judgments for each item.

A total of 130 crowdworkers (83 females; average age = 31 years) were recruited to take part in either of the studies. All participants were native English speakers residing in English-speaking countries and registered on Prolific as participant. In Study 1 (TED-MDB), the experiment took 15 minutes on average, and participants were remunerated with 1.25 GBP. We anticipated that Study 2 would take longer because it required more reading, we therefore remunerated it with 2.50 GBP, but it turned out to only take 17 minutes on average.

4 Language Model

Connectives were compared with respect to the average negative log likelihood of the sentence including the connective, using OpenAI's pretrained Generatively Pre-trained Transformer (GPT-2) language model (Radford et al., 2019) implemented in the Transformers library of huggingface (Wolf et al., 2020). GPT-2 is a unidirectional transformer-based language model trained on a dataset of 40 GB of web crawling data.

The model compared the same set of connective choices as given to the human crowdworkers. The connective choice resulting in the best sequence according to GPT-2 was selected as the predicted connective.

	Agreement		
Data	majority	random	most common
TED-MDB	.723	.223	.852
PDTB-3	.454	.242	.304

Table 1: Agreement of the majority, random and most common connective choices with the original connective.

5 Results

5.1 Agreement between original connective, crowd choices and simple heuristic

Table 1 compares, per corpus, the original connective to three connective choices. The first column presents the agreement between the original connective and the majority choice; that is, the connective most frequently chosen by the workers for each instance. The second column presents the agreement with a randomly chosen connective; this was based on a baseline where a connective was chosen at random from the list of connectives that can express that relation type. Finally, the third column presents the agreement with the most common connective: a simple heuristic was used to select the connective that was most frequently chosen by workers for a discourse relation type, i.e. collapsing across the different instances of a discourse relation type. This last metric measures whether crowd workers use a default connective for every relation sense (e.g., always use *and* to express a conjunction, irrespective of the original connective that may have expressed that relation.)

We see large differences between the studies: In the TED-MDB dataset, which represents the natural distribution of the types of discourse relation, agreement between the majority choice among

workers and the original connective is relatively high (72.3%): most workers chose the connective that matches the original connective in the corpus (mostly *and*, *but*, *because*, *so*). The agreement between the original connective and the "most common" baseline is even higher (85.2%), implying that the original connective of an instance is usually the most common connective for the particular relation type. This high agreement is due to the skewed distribution of the relations in TED-MDB: out of the 210 samples, there are 78 instances of *and* for CONJUNCTION and 34 instances of *but* for ARG2-AS-DENIER or CONTRAST (see Table 3), and these connectives are the most common connectives for these relation types. Such a skew towards a small number of highly frequent connectives is particularly strong in spoken language, while a larger variety of connectives is usually observed in written domains (Crible and Cuenca, 2017).

We conclude that for the spoken domain, simply generating the most frequent connective for a given relation sense already yields quite high accuracy. However, from this analysis, we cannot assess what happens in the cases where a less frequent relation needs to be expressed, as the number of these cases in the TED-MDB set is too small for reliable analysis, and we cannot tell whether this result would transfer to the written domain. Therefore, we next conducted an analysis using a balanced dataset based on the PDTB-3.

As Table 1 shows, the agreement is a lot lower for the PDTB-3-based balanced dataset. Here, the majority choice among crowdworkers agreed with the original connective for 45% of the items, while the most frequently selected connective of a relation type is identical with the original connective in only 30% of cases (this makes sense, as it reflects how the dataset was assembled). We will analyse these cases in more detail in the next section.

5.2 Analysis per relation type – PDTB-3

Table 2 shows that the accuracy for humans in recovering the original connective differs quite a lot depending on the relation and the original DC (ranging from 15% to 67%). Table 4, in the Appendix, presents more results on the agreement per relation type (collapsing across the original connectives) and statistics to test whether there is a significant difference between the human label distributions among items of different original connectives but the same relation sense.

In the remainder of this section, we present a more fine-grained analysis, for which we classified the results per relation sense and connective into the following broad categories: 1) Freely interchangeable connectives – humans don't show a preference for the original connective, several connectives are equally preferred. 2) Human preference is in line with original. 3) Human preference is for a specific connective which is different from the original.

Case 1: Freely interchangeable connectives We observe some relations for which participants chose a wide variety of connectives, and where their choice does not necessarily agree with the original connective. Consider, for example, the relation sense `Arg2-as-denier`: here, the distribution of insertions is nearly identical for all of the original connectives. These cases are unproblematic from a generation perspective, as any choice would be acceptable. Note though that the language model gives higher probability to the more generic connective *but*, while humans have a slight preference for the more specific connective *however*.

We also observe a high rate of interchangeability between *for example* and *for instance*, which mark the relation `Arg2-as-instance`, as well as the connectives for the `Arg2-as-detail` relation. In these cases, the distributions of the human choices are not significantly different from a uniform distribution.

For `Result` relations, the connectives *thus*, *as a result* and *therefore* seem to be freely interchangeable, while *and* and *so* have slightly different distributions. Interestingly, the language model seems to be able to pick up on cues that indicate the connective *so*. We note, though, that it incorrectly assigns the connective *and* to relations which were originally marked by *so* or *thus*, and for which *and* is not a preferred option according to human insertions – the language model here prefers a connective which is too generic.

Case 2: Crowd-workers agree with the original An interesting case are instances for which the humans prefer a connective that strongly matches with the original connective, while not selecting alternatives from the set of connectives for the same relation. In these cases, the results suggest that the relation sense is not sufficiently detailed, and there are relevant aspects of meaning of the relation which is not captured by the relational label.

Among these cases, we can then further distin-

Relations (count)	Original connectives (count)	Distribution of humans' choices					Most frequent LM choice
		but	*however*	*though*	*still*	*yet*	
Arg2-as-denier	**but (4)**	23%	**33%**	10%	5%	30%	**but**
(20)	**however (4)**	23%	**30%**	18%	15%	15%	**but**
Case 1	**though (4)**	28%	**38%**	15%	8%	13%	**but**
	still (4)	25%	**35%**	23%	5%	13%	**but**
	yet (4)	23%	**35%**	13%	10%	20%	**but**
		for example	*for instance*	*in particular*	*in fact*	*as*	
Arg2-as-inst.							
(18)	**for example (9)*****	30%	**41%**	18%	1%	10%	**for example**
Case 1	**for instance (9)**	**29%**	20%	18%	17%	17%	**for example**
		so	*thus*	*as a result*	*therefore*	*and*	
Result	**so (6)****	**40%**	13%	17%	28%	2%	**so**
(18)	**thus (6)**	8%	23%	**33%**	25%	10%	**and**
Case 1	**as a result (6)**	10%	25%	**32%**	15%	18%	**and, as a result**
		indeed	*in fact*	*specifically*	*in particular*	*and*	
Arg2-as-detail	**indeed (4)***	13%	**53%**	5%	13%	18%	**and, in fact**
(16)	**in fact (4)**	5%	25%	25%	15%	**30%**	**and**
Case 1/3	**specifically (4)**	13%	23%	**25%**	18%	23%	**and**
	in particular (4)	8%	13%	20%	25%	**35%**	**and**
		although	*though*	*even if*	*even though*	*while*	
Arg1-as-denier	**although (4)****	23%	8%	3%	**60%**	8%	**even though**
(16)	**though (4)***	35%	8%	8%	**43%**	8%	**even though**
Case 2A	**even if (3)**	10%	10%	**50%**	23%	7%	**even if**
	even though (5)***	10%	0%	18%	**66%**	6%	**even though**
		then	*before*	*until*	*later*	*ultimately*	
Precedence	**then (5)**	28%	8%	26%	8%	**30%**	**then**
(14)	**before (5)*****	0%	**84%**	14%	2%	0%	**before**
Case 2A	**until (4)*****	3%	8%	**90%**	0%	0%	**until**
		because	*as*	*since*	*for*	*when*	
Reason	**because (6)*****	**38%**	35%	25%	0%	2%	**because**
(18)	**as (6)***	25%	**32%**	17%	2%	25%	**as**
Case 2A	**since (6)*****	22%	32%	**38%**	0%	8%	**as**
		and	*also*	*in addition*	*moreover*	*or*	
Conjunction	**and (6)*****	**57%**	3%	20%	17%	3%	**and**
(19)	**also (6)*****	**53%**	10%	18%	17%	2%	**and**
Case 2A/B	**in addition (7)*****	**41%**	3%	27%	29%	0%	**and**
		after	*when*	*since*	*once*	*previously*	
Succession	**after (4)****	**48%**	15%	5%	33%	0%	**after, when**
(17)	**when (4)****	13%	**48%**	5%	35%	0%	**once**
Case 2B	**since (3)*****	7%	7%	**83%**	3%	0%	**since**
	once (3)***	10%	33%	0%	**57%**	0%	**once**
	previously (3)	**40%**	3%	17%	10%	30%	**when**
		but	*while*	*however*	*still*	*on the other hand*	
Contrast	**but (4)****	**40%**	8%	40%	5%	8%	**but**
(15)	**while (4)****	10%	**60%**	18%	3%	10%	**but, while**
Case 2B	**however (3)***	13%	20%	**57%**	7%	3%	**but**
	still (4)	**38%**	10%	25%	3%	25%	**but**
Arg2-as-cond.		*if*	*when*	*until*	*as long as*	*in case*	
(18)	**if (9)*****	30%	16%	2%	**43%**	9%	**as long as**
Case 2B	**when (9)*****	16%	**49%**	9%	19%	8%	**when**
		when	*as*	*while*	*meanwhile*	*at the same time*	
Synchronous	**when (4)***	**45%**	20%	23%	3%	10%	**when**
(18)	**as (4)*****	15%	**55%**	25%	0%	5%	**as**
Case 2B	**while (3)*****	13%	8%	**80%**	0%	0%	**at the same time**
	meanwhile (3)	3%	13%	30%	**43%**	10%	**at the same time**
	at the same time (3)	7%	27%	**30%**	20%	17%	**at the same time**

Table 2: Distribution of connectives per relation sense chosen by crowd workers in the PDTB data. The relation senses are ordered according to their case classification in section 5.2. Under the column "Distribution of humans' choices", the majority choice per original connective type is in marked in bold, and the choice matching with the original connective is underlined. χ^2 **tests** are performed to test whether the distributions are significantly different from a uniform distribution (*:$p < 0.05$;**:$p < 0.01$;***:$p < 0.001$).The final column presents the language model's (GPT2) prediction.

guish between cases where a language model successfully picks up on the distinction (Case 2A) vs. cases where the language model assigns a dispreferred connective (Case 2B).

Examples of cases of type 2A include the relation `Arg1-as-denier`: here, the crowd workers had a strong preference for the connective *even though*, which they inserted for sentences originally marked with *although*, *though* and *even though*. These connectives seem to be fully interchangeable, and could even be classified as Case 1, if it wasn't for the connective *even if*, which shows a different distribution. Interestingly, the language model agrees with human preferences for these cases – it generally assigns highest likelihood to the connective *even though*, but also correctly marks the original cases that were marked with *even if*.

A similar case is `Precedence`: the connectives *then*, *before* and *until* can be successfully recovered based on the language model.

Next, we take a look at the `Reason` relation. Here, participants largely recover the original connective, indicating that these connectives are not fully interchangeable. In particular, the connective *because* places more emphasis on the reason, and so it should be used in cases where the information given as the reason is new. *As* and *since* place emphasis on the result and are more commonly used when the speaker believes that the content of the reason is something that the addressee already knows. Interestingly, the language model successfully recovers the difference between connectives *because* and *as*, but prefers *as* also for cases where the original connective is *since*. This confusion might be related to both connectives being able to express a purely temporal relation as well. We can also see that only those `Reason` relations originally marked by *as* can be marked using *when*, which indicates that relations marked with *as* have a stronger temporal aspect than ones marked with *because*. A model that captures this aspect would have to learn when a reason is not temporal.

Finally, for `Conjunction` relations, we observe that both humans and the language model prefer to simply use the connective *and*. A small difference in human insertions is present between the instances containing the connective *in addition* originally vs. *and* or *also*: the heavier connective *in addition* can be more easily replaced by *moreover* compared to lighter connectives. A model using these connectives correctly would hence have to learn to pick up on heaviness effects relating to the prominence and length of relational arguments.

Examples of Case 2B are found in `Succession` relations. Here, the connectives *after*, *since* and *once* are successfully recovered by the workers as well as the language model, but for instances originally expressed by the connective *when*, the language model prefers *once*. A closer look at the data for the connective *when* reveals that in two cases, *once* is also the preferred connective for humans. For the other two items, *once* is not used at all. We find that this difference is related to tense/time frame: *Once* is interchangeable with *when* when the items speaks of future possibilities, not about the past. To illustrate, compare (1), which can be expressed with both *once* and *when*, with (2), can only be expressed with *when*.

(1) (...) [the plant can be reactivated quickly] ___ [the market improves.]

(2) [The controls on cooperatives appeared relatively liberal] ___ [first introduced.]

For `Succession` instances originally marked with *previously*, the language model assigns the connective *when*. However, humans only inserted *when* in 3% of cases. Manual analysis shows that this is because *previously* is used for a change in state/event, as in (3). The connective *when* is dispreferred in such cases, because it implies a shorter time frame.

(3) [Equus Capital Corp. would pay $12 cash for each of Tony Lama's 2.1 million shares outstanding] ___ [it offered $13.65 a share in cash, or $29 million.]

In `Contrast` relations, the human participants mostly recovered the original connective, indicating that they are not freely interchangeable. However, the language model uniformly selects *but* as the connective with highest likelihood. This seems to be particularly problematic for the instances originally marked with *while* and *however*, as they exhibit a low rate of *but*-choices among the humans. An example item for a case originally marked with *while* is shown in (4). An important factor here seems to be related to the simultaneous nature of the two facts that are contrasted.

(4) [Among liberals, 60% have positive views of her] ___ [50% approve of the president's job performance.]

For items originally marked with *however*, *but* also seems acceptable , but *however* gives a slightly stronger marking and breaks up the two arguments into two separate sentences. Information-theoretic aspects might be at play here. To illustrate this, consider the following example:

(5) One of the fastest growing segments of the wine market is the category of super-premiums (...). [For years, this group included a stable of classics – Bordeaux first growths (Lafite-Rothschild, Latour, Haut-Brion, Petrus), Grand Cru Burgundies (Romanee-Conti and La Tache) deluxe Champagnes (Dom Perignon or Roederer Cristal), rarefied sweet wines (Chateau Yquem or Trockenbeerenauslesen Rieslings from Germany, and Biondi-Santi Brunello Riserva from Tuscany)] ___ [in the last year or so, this exclusive club has taken in a host of flashy new members.]

But would be felicitous in this example, yet only one participant selected *but*, and eight selected *however*. Note that the first argument of this relation is rather long, and so participants might prefer however because it provides a clearer break between the arguments: in natural language, *however* is more frequently used to start a new sentence than *but*, which is more frequently used sentence-medially.

In `Arg2-as-condition` relations, we observe a low prediction accuracy because workers avoid the strong connective *if* – they might have interpreted multiple relations (and thus prefer multi-sense DCs over single-sense DCs). Instead of *if*, most participants and the language model chose *when* or *as long as*. These are however not suitable for all usages of *if*, as there may be differences in whether an event will actually happen, or whether there is uncertainty about it happening at all. Consider (6) and (7). Both were originally marked with *if*, but *as long as* was the preferred choice for (6). For (7), however, *as long as* would be infelicitous, and *if* was indeed the preferred crowd-choice.

(6) [The IRS says people in the disaster areas won't be penalized for late filing] ___ [their returns are marked "Hugo" and postmarked by Jan. 16.]

(7) [What will Mr. Sagan do with his new theater building] ___ [the allure of Hollywood and Broadway proves too much for such Steppenwolf stalwarts as John Malkovich (...).]

Finally, for the temporal `Synchronous` relations, the language model prefers the connective *at the same time* for relations marked originally by *at the same time*, *while* or *meanwhile*. By contrast, the workers agreed with the original connectives for both *while* and *meanwhile* items. For *at the same time* items, the workers also preferred *while*. A closer look at the instances reveals that *while* works particularly well when there is also a possible contrastive aspect in the item, as in (8).

(8) [Personal spending grew 0.2% in September to a $3.526 trillion annual rate] ___ [personal income was held down by the effects of Hurricane Hugo.]

Case 3: Humans disagree with original We found disagreements between humans and original connectives mostly in cases where the original connective was relatively unusual in sentence-initial position, such as *yet*, *also*, *still* or *indeed*. It is possible that this results from our experimental design, as the connective options were presented without a comma, which would be needed for many of these cases if used in sentence-initial position.

6 Discussion

We set out to investigate the extent to which the choice of connective to realize a particular relation is constrained. Specifically, we saw three possible outcomes: connectives that express a relation are fully interchangeable for the PDTB-3 relations that we work with here, other factors (e.g., information-theoretic ones) lead to preferences even though interchangeability is given, or the relation sense distinctions are not sufficiently fine-grained, so that an inappropriate connective might be chosen.

The main result is that connective choice varies depending on the dataset: in TED-MDB, high accuracy is achieved by choosing the most probable realization according to GPT-2 or choosing the most common connective per relation type. In PDTB, accuracy of the connective choice model is a lot lower.

We identified a set of different cases: a very simple connective selection strategy of randomly choosing among fitting connectives, might be good enough for some relations (Case 1). For other relations (identified as Case 2A above), a language-model-based strategy seems to perform well.

However, we also identified some more problematic cases (classified as 2B above): for these

cases, we would either need to extend the relation inventory in order to capture the more fine-grained distinctions, such as temporal aspects (in `Succession` relations), factuality of events (in `Arg2-as-condition`) or whether there is a shared common ground (`Reason` relations). Alternatively, a more sophisticated language model would have to be developed, which can learn to use the correct connective in these cases. In future work, we aim to evaluate to what extent this could be learned by a transformer model which has been fine-tuned on a connective insertion task where the arguments and the target relation is given, and the correct connective needs to be selected.

We also found some limited evidence for other factors influencing connective choice – for the selection of a rather common and light connective like *and* or *but* vs. a more heavy and longer connective like *in addition* or *on the other hand*, it seems that while substitutability is given, preferences in terms of focus on the relation and heaviness and distance of the relational arguments may play an important role in connective choice. Again, it would be interesting to see whether a more sophisticated language model could pick up on these aspects.

The analyses performed here are clearly empirical. We do not claim that our method has allowed us to detect all possible cases where substitutability of connectives might not be given. Instead, our analyses provide a practical perspective indicating how a simple generation system for connectives would fare for frequent relations and connectives.

In relation to this, the difference in performance of the generation system on TED-MDB and PDTB provides interesting insights. These datasets represent different domains: the TED-MDB dataset consists of spoken data, whereas the PDTB dataset consists of written data. Spoken data tends to be characterized by a smaller range of connective types, as was the case in our data. The model's performance on this dataset seems promising, as it generally selects the same connective as the original one. As our subsequent analysis shows, this is mostly because these relations are "easy" – the original connective in TED-MDB is often the most common connective for a given relation type (e.g., *and*, *but*).

In the PDTB dataset, we manipulated the range and distribution of connectives by specifically choosing instances of a larger number of connectives, to give a more representative picture of the limitations of the simple GPT-2 based model. Some of the connectives in this set are infrequent in natural language. We find that the GPT-2 tends to default to frequently occurring connectives within relation senses (e.g., *but*, *and*, *even though*), irrespective of the original connective.

Finally, we note that the results of our analysis of course also depend on our choice of using PDTB as a framework here – other frameworks can differ in what coherence relations they distinguish, and accordingly, the results regarding which relations need to be distinguished at a more fine-grained level might vary. The methodology used here can also be applied to data from other frameworks to evaluate whether those distinctions are sufficient for computational purposes such as connective generation.

7 Conclusion

The current study showed that, in spoken data, a language generation system can predict the connective that should be used to express a discourse relation with high accuracy. This is partially because the relations tend to be quite simple and be marked by high-frequency connectives: even simply choosing the most prototypical connective for a given relation sense shows high accuracy on this data set.

On the other hand, our subsequent analysis of written data, using a more balanced set of relations and connectives, highlighted that there is a clear need for a more sophisticated method for connective choice. The results from human annotators indicated that connectives for many relational classes are not fully interchangeable. In some cases (`Succession`, `Condition`, and `Reason` relations), additional finer-grained relation types are needed to capture more information. In other cases (relating to specific connectives, such as *in addition* and *however*), information-theoretic constraints appear to influence connective choice. These insights can be useful for natural language generation researchers as well as research on automatic discourse parsing.

Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding".

References

Ludivine Crible and Maria-Josep Cuenca. 2017. Discourse markers in speech: characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166.

Alistair Knott. 1996. *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, The University of Edinburgh: College of Science and Engineering: School of Informatics.

Alistair Knott, Ted Sanders, and Jon Oberlander. 2002. Levels of representation in discourse relations.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from gpt-2. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 52–59.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. *arXiv preprint arXiv:2106.03192*.

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic prediction of discourse connectives. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. *arXiv preprint arXiv:1704.00217*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

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.

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. IEEE.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.

Zhi Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Coling 2010: Posters*, pages 1507–1514.

A Appendix

Relations (count)	Original connectives (count)	Options of alternative connectives
Conjunction (80)	and (78), also (2)	and, also, in addition, furthermore, (nothing)
Disjunction (3)	or (3)	or, alternatively, otherwise
Arg1-as-detail (1)	in short (1)	in short, in summary, to sum up, in general, to conclude
Arg2-as-detail (4)	in fact (2), clearly (1), especially (1)	in fact, clearly, especially, in more detail, in particular
Arg2-as-manner (6)	by (4), through (2)	by, through, by means of, by way of
Reason (17)	because (16), since (1)	because, since, as
Result (17)	so (15), because of that (1), that's why (1)	so, because of that, that's why, therefore, as a result
Arg2-as-condition (20)	if (17), when (3)	if, when, provided that, given that, in case
Contrast (15)	but (10), and (2), when (1), where (1), where, on the one hand (1)	and, but, when, on the contrary, on the one/other hand
Arg2-as-denier (26)	but (24), however (1), though (1)	but, however, though, nonetheless, despite this
Similarity (1)	also (1)	also, similarily, in the same vein
Precedence (5)	and then (4), then (1)	and then, then, and next, and afterwards
Synchronous (15)	as (7), when (6), while (1), at the same time (1)	as, when, at the same time, while, meanwhile

Table 3: Distribution of discourse relation and connective types of the items from TED-MDB and the choices of alternative connectives given to the human crowdworkers

Relation types (item count)	Original connectives (item count) / connective options	Agree with org. Maj.	LM	χ^2 test χ^2	*df*	
Conjunction (19)	and (6), also (6), in addition (7), *moreover, or*	.316	.632	12.01	8	
Arg2-as-detail (16)	indeed (4), in fact (4), specifically (4) in particular (4), *and*	.200	.067	24.81	12	*
Arg2-as-instance (18)	for example (9), for instance (9), *in particular, in fact, as*	.278	.389	20.33	4	***
Reason (18)	because (6), as (6), since (6), *for, when*	.444	.555	20.78	8	**
Result (18)	so (6), thus (6), as a result (6), *therefore, and*	.444	.500	36.51	8	***
Arg2-as-condition (18)	if (9), when (9), *until, as long as, in case*	.444	.278	31.95	4	***
Arg1-as-denier (16)	although (4), though (4), even if (3), even though (5), *while*	.500	.500	46.82	12	***
Arg2-as-denier (20)	but (4), however (4), though (4), still (4), yet (4)	.150	.250	11.23	16	
Contrast (15)	but (4), while (4), however (3), still (4), *on the other hand*	.530	.400	58.30	12	***
Precedence (14)	then (5), before (5), until (4), *later*, *ultimately*	.786	.929	142.60	8	***
Succession (17)	after (4), when (4), since (3), once (3), *previously*	.611	.500	179.72	16	***
Synchronous (18)	when (4), as (4), while (3), meanwhile (3), at the same time (3)	.611	.389	116.84	16	***
Overall (215)		.423	.433			

Table 4: Distribution of discourse relation and connective types of the experimental items from PDTB-3. The **connective options** given to the human crowdworkers primarily include the original connectives of the same relation in the sample set. Further options (in *italics* and without item counts) are added such that there are five choices per question. **Agree with org.** is the percentage of items per relation where the majority choice by the crowdworkers (**Maj.**) or the choice with the lowest perplexity based on GPT-2 (**LM**) matches the original connective. Here, the χ^2 **test** is performed to test if there is significant difference between the human label distributions among items of different original connectives but the same relation (*:$p < 0.05$;**:$p < 0.01$;***:$p < 0.001$).