SciPara: A New Dataset for Investigating Paragraph Discourse Structure in Scientific Papers

Anna Kiepura[†], **Yingqiang Gao**[†], **Jessica Lam**[†], **Nianlong Gu**[‡]
Richard H.R. Hahnloser[†]
[†]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland
{akiepura, yingqiang.gao, lamjessica, rich}@ini.ethz.ch
[‡]Linguistic Research Infrastructure, University of Zurich, Switzerland
nianlong.gu@uzh.ch

Abstract

Good scientific writing makes use of specific sentence and paragraph structures, providing a rich platform for discourse analysis and developing tools to enhance text readability. In this vein, we introduce SciPara[1], a novel dataset consisting of 981 scientific paragraphs annotated by experts in terms of sentence discourse types and topic information. On this dataset, we explored two tasks: 1) discourse category classification, which is to predict the discourse category of a sentence by using its paragraph and surrounding paragraphs as context, and 2) discourse sentence generation, which is to generate a sentence of a certain discourse category by using various contexts as input. We found that Pre-trained Language Models (PLMs) can accurately identify Topic Sentences in SciPara, but have difficulty distinguishing Concluding, Transition, and Supporting Sentences. The quality of the sentences generated by all investigated PLMs improved with amount of context, regardless of discourse category. However, not all contexts were equally influential. Contrary to common assumptions about well-crafted scientific paragraphs, our analysis revealed that paradoxically, paragraphs with complete discourse structures are less readable.

1 Introduction

Writing a scientific paper that is understandable to readers is a challenging task. Well-written scientific papers not only facilitate the comprehension of scientific discoveries but also reduce the risk of disseminating inaccuracies and misconceptions in research (Freeling et al., 2021).

As a rhetorical unit of writing, paragraphs contain valuable information regarding the logical and narrative connections among sentences (Nunan, 2015). Scientific papers with many well-written paragraphs are easier to understand. In those paragraphs, related sentences are grouped and information is stitched in a thematically progressing manner (Weissberg, 1984).

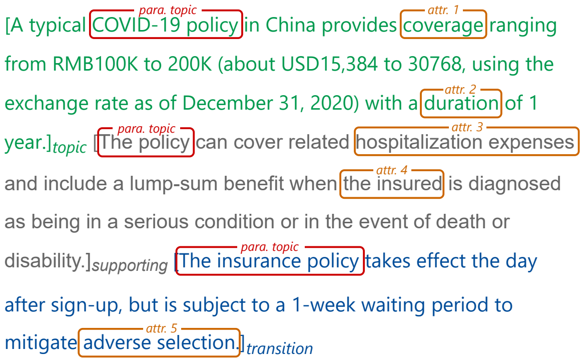

[A typical COVID-19 policy in China provides coverage ranging from RMB100K to 200K (about USD15,384 to 30768, using the exchange rate as of December 31, 2020) with a duration of 1 year.]$_{topic}$ [The policy can cover related hospitalization expenses and include a lump-sum benefit when the insured is diagnosed as being in a serious condition or in the event of death or disability.]$_{supporting}$ [The insurance policy takes effect the day after sign-up, but is subject to a 1-week waiting period to mitigate adverse selection.]$_{transition}$

Figure 1: An example (taken from Feng et al. (2023)) annotated paragraph with one Topic Sentence (green), one Supporting Sentence (grey), and one Transition Sentence (blue). The paragraph topic is indicated in red and the topic attributes are indicated in orange.

In recent years, significant efforts have been directed at utilizing NLP technologies to process and comprehend scientific texts. For instance, research has focused on automatic summarization (Gu et al., 2022), text generation (Hu and Wan, 2014; Wang et al., 2019; Chen et al., 2021), as well as argument mining and discourse analysis (Fergadis et al., 2021; Gao et al., 2022; Achakulvisut et al., 2019), all in the context of scientific papers. However, few efforts have been devoted to identifying well-written scientific paragraphs from the perspective of discourse structure.

In this work, we propose **Sci**entific **Para**graphs (**SciPara**), a novel dataset specifically curated for studying the structure of scientific paragraphs. SciPara is a collection of scientific paragraphs that have been manually annotated by professional editors with strong biomedical backgrounds. The annotations include paragraph-level discourse com-

[1] Code and data are available at https://github.com/annamkiepura/SciPara.

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 12–26
March 21, 2024 ©2024 Association for Computational Linguistics

pleteness, sentence-level discourse categories, and word-level occurrences of the paragraph topic. By training various language models on SciPara, we address the following research questions (RQs):

RQ1 Can language models distinguish sentences of different discourse categories?

RQ2 Can Topic, Concluding, and Transition Sentences be generated from the rest of the corresponding paragraphs?

RQ3 Are paragraphs with complete discourse structure more readable?

Our main **contributions** are as follows: 1) We propose a manually annotated dataset of scientific paragraphs, which is, to the best of our knowledge, the first dataset specifically designed for the study of the discourse structure of scientific paragraphs; 2) We fine-tune language models to perform sentence classification and generation tasks on our dataset; 3) We perform an in-depth analysis of the paragraph discourse structure with respect to our experimental results.

2 SciPara: A New Dataset for Discourse Structure of Scientific Paragraphs

Our goal is to facilitate the analysis of scientific paragraph discourse structure on two levels:

Sentence level How do individual sentences relate to the paragraph's discourse structure?

Subsentence level What are the paragraph's topic and its corresponding attributes?

In this section, we outline the protocol given to the annotators for creating SciPara (see Figure 2a).

2.1 Initial paragraph filtering

To preserve the coherence of the paper's narrative, annotators processed paragraphs in their order of occurrence. We instructed annotators to skip paragraphs that had parsing errors, such as incorrect sentence splits, or that contained less than three sentences. The annotators were required to label such paragraphs as "Bad Parse" and "Too Short" respectively.

2.2 Sentence-level annotation

We tasked annotators with categorizing each sentence of a paragraph into one of the following six discourse categories:

Topic Sentence A sentence that encapsulates the central theme of the paragraph. The information presented in a Topic Sentence is typically expanded upon in the other sentences of the paragraph (McCarthy et al., 2008).

Supporting Sentence A sentence that bolsters the Topic Sentence(s) with relevant information such as explanations, elaborations, and examples.

Concluding Sentence A sentence that summarizes and closes the narrative of the paragraph.

Transition Sentence A sentence that connects the current paragraph to the next paragraph, thereby maintaining the coherence of the paper.

Off-Topic Sentence A sentence that lacks information pertinent to the topic of the paragraph.

Redundant Sentence A sentence whose content has already been stated in an earlier sentence of the paragraph.

We refer to paragraphs with at least one Topic Sentence and at least one Concluding or Transition Sentence as paragraphs with *complete* discourse structure. All other paragraphs are considered to have an *incomplete* discourse structure (see Table 2).

During the annotation, a few paragraphs turned out to have no Topic Sentences. We instructed annotators to halt the annotation of such paragraphs and to proceed to the next.

2.3 Subsentence-level annotation

The annotators then moved on to the subsentence-level task, see Figure 2b. The first step was to identify noun phrases in the Topic Sentence(s) that pertained to the topic of the paragraph. Inspired by Ajjour et al. (2023), we defined the paragraph topic hierarchically:

Topic Ontology A noun phrase that best encapsulates the topic of the paragraph.

Topic Attribute A noun phrase that describes an aspect of the Topic Ontology.

We allowed for exactly one Topic Ontology and up to seven unique Topic Attributes per paragraph. A handful of paragraphs had multiple Topic Sentences; however, in all cases, the multiple Topic Sentences had the same Topic Ontology and Topic Attributes.

Next, the annotators identified all re-occurrences of the paragraph topic (both Topic Ontology and

Topic Attributes) in the other sentences of the paragraph. Re-occurrences could be either exact matches or semantically similar noun phrases.

When a sentence s did not contain the paragraph topic, we asked annotators to identify links between s and the other sentences of the paragraph. Links are noun phrases that can be found in both s and at least one other sentence of the paragraph that contains the paragraph topic annotations. Finally, we asked annotators to label sentences that contain neither the paragraph topic nor links as "Off-Topic Sentences".

2.4 Data sources

We obtained 62 scientific papers from two datasets: Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020) and Europe PMC[2]. S2ORC is a comprehensive repository consisting of 81 million scientific papers in English. Europe PMC is an open-access repository containing 43 million publications and preprints enriched with links to supporting data, reviews, and other relevant sources.

We investigated paragraphs from INTRODUCTION and DISCUSSION sections only. This is because these sections aim to deliver narratives, as compared to, say, RESULTS sections, which typically aim to list but not necessarily analyse the papers' findings (Nair et al., 2014).

Due to the need for clear sectioning, we only used papers from the fields of medicine and biomedicine. Papers from such fields often follow the IMRaD format and contain INTRODUCTION, METHODS, RESULTS, and DISCUSSION sections (Nair et al., 2014).

For the annotation tasks, we enlisted the expertise of four proficient biomedical editors who are members of the European Medical Writers Association (EMWA)[3]. Annotation was performed on the interactive data annotation platform *Doccano* (Nakayama et al., 2018).

2.5 SciPara statistics

The SciPara dataset consists of 981 paragraphs and 4071 sentences, see Table 1. Across these paragraphs, the annotators identified more than 700 instances of Topic Ontologies and over 2800 instances of Topic Attributes. In total, 432 paragraphs have complete discourse structure and 309 paragraphs have incomplete discourse structure, see Table 2. We kept the 240 paragraphs that were not annotated for discourse completeness so that we could study the influence of context information in the discourse sentence generation task.

Statistic	Count	Statistic	Count
# Papers	62	# Topic Sentences	724
# Paragraphs	981	# Supporting Sentences	2,869
# Sentences	4,071	# Concluding Sentences	273
# Topic Attribute	2,821	# Transition Sentences	188
# Topic Ontology	724	# Off-topic Sentences	3
-	-	# Redundant Sentences	14

Table 1: Overall statistics of our SciPara dataset.

Topic Sentence	Concluding Sentence	Transition Sentence	Discourse Structure	Count
✓	✓	✗	Complete	250
✓	✗	✓	Complete	177
✓	✓	✓	Complete	5
✓	✗	✗	Incomplete	284
✗	✓	✗	Incomplete	7
✗	✗	✓	Incomplete	4
✗	✗	✗	Incomplete	14

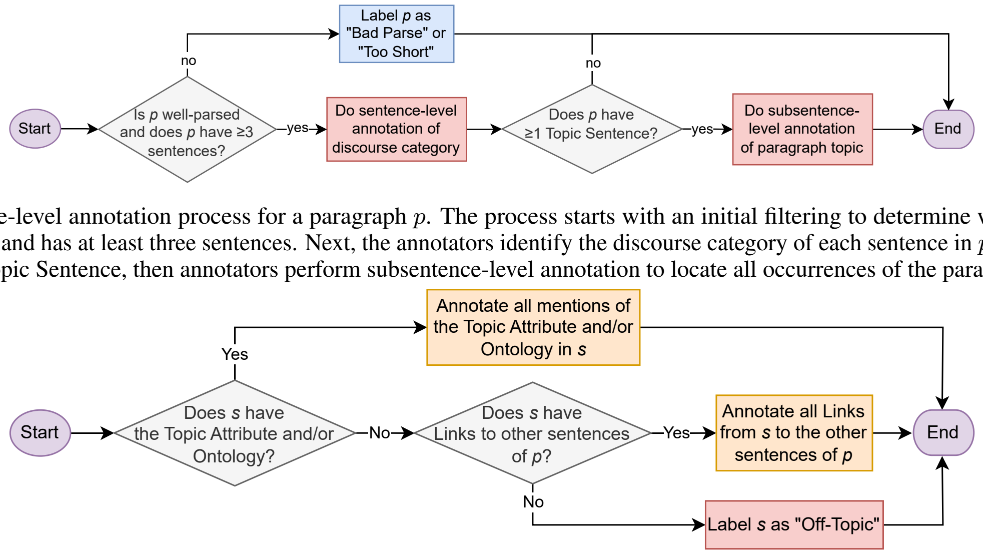
Table 2: Structure assessment for sentence-level annotation. We exclude paragraphs with both Concluding and Transition Sentences but no Topic Sentences on purpose, since subsentence-level annotation for this type of paragraphs was not possible (Topic Ontology must be labeled from the Topic Sentence).

Due to the unexpected absence of annotator 3, we present the inter-annotator agreement (IAA) results for annotators 1, 2, and 4 only, see Table 3. For sentence-level annotation, we calculated Cohen's Kappa coefficients (Cohen, 1960) for each pair of annotators. As for subsentence-level annotation, where Topic Ontology and Topic Attributes do not have fixed discourse categories and can vary in length, we evaluated the IAA based on lexical overlap of annotations measured by ROUGE scores (Lin, 2004). High ROUGE-1 and ROUGE-2 scores therefore indicate better agreement between pairs of annotators.

Sentence-level annotations of Topic, Supporting, and Concluding discourse categories showed a high agreement among annotators when compared against a reference rubric for Cohen's Kappa scores interpretation (McHugh, 2012), which we summarize in the legend of Table 3. For example, for Topic Sentence identification, all of the analyzed data subsets fall into the "strong agreement" category. This indicates that the task of identifying these discourse types was clearly defined and the annotators understood the instructions well. The

[2] https://europepmc.org/

[3] https://www.emwa.org


Label *p* as "Bad Parse" or "Too Short"
no
no
Start
Is *p* well-parsed and does *p* have ≥3 sentences?
yes
Do sentence-level annotation of discourse category
Does *p* have ≥1 Topic Sentence?
yes
Do subsentence-level annotation of paragraph topic
End
Annotate all mentions of the Topic Attribute and/or Ontology in *s*
Yes
Start
Does *s* have the Topic Attribute and/or Ontology?
No
Does *s* have Links to other sentences of *p*?
Yes
Annotate all Links from *s* to the other sentences of *p*
End
No
Label *s* as "Off-Topic"

(a) Sentence-level annotation process for a paragraph p. The process starts with an initial filtering to determine whether p is well-parsed and has at least three sentences. Next, the annotators identify the discourse category of each sentence in p. If p has at least one Topic Sentence, then annotators perform subsentence-level annotation to locate all occurrences of the paragraph topic.

(b) Subsentence-level annotation process for a paragraph p. Starting with labeling the Topic Ontology in the Topic Sentence s, the subsentence-level annotation identifies Topic Attributes throughout the paragraph.

Figure 2: Overview of the SciPara data annotation process for a given paragraph p.

Subset	A			B	C
Annotator Group	1&2	2&4	1&4	1&4	1&4
κ - Topic Sent.	0.79	0.72	0.92	**0.97**	0.96
κ - Supp. Sent.	0.75	0.68	**0.78**	0.68	**0.78**
κ - Concl. Sent.	0.69	**0.78**	0.60	0.64	0.53
κ - Trans. Sent.	**0.69**	0.40	0.22	0.08	0.26

(a) IAA results for sentence-level annotation.

Subset	A			B	C
Annotator Group	1&2	2&4	1&4	1&4	1&4
R-1 (f-measure)	0.59	0.61	**0.67**	0.63	0.61
R-2 (f-measure)	0.43	0.48	**0.50**	0.45	0.41

(b) IAA results for subsentence-level annotation (stopwords are removed for all measures).

Table 3: Inter-annotator agreement (IAA) results for sentence-level and subsentence-level ($\kappa \leq 0.4$ = poor agreement; $0.4 < \kappa \leq 0.6$ = fair agreement; $\kappa > 0.6$ = strong agreement). Subsets A, B, and C contain 36, 36, and 50 paragraphs, respectively.

agreement was considerably lower for Transition Sentences, which we discuss in more detail in Limitations.

For subsentence-level annotations, given that the average length of Topic Ontology and Topic Attributes was around 3 to 4 words, a lexical overlap score above 0.4 is considered as high. Thus, it suggests that the subsentence-level task was also well understood by the annotators, suggesting that the curated dataset has good quality.

3 Methods

In the following section, we detail the experimental methods applied to the SciPara dataset to address our research questions. Notably, our experiments primarily utilized the annotations corresponding to the sentence-level task, and the Topic Ontology annotations from the subsentence-level task. We plan to incorporate other annotation types, such as Topic Attributes and links, in future studies.

3.1 Discourse category classification

Underlying RQ1 is the following sequential sentence classification task (Cohan et al., 2019): Identifying the discourse category of a sentence Y based on the context of Y. By context, we refer to the paragraph P containing Y and the subsequent paragraph P'. We ignored Off-Topic and Redundant Sentences because of their rarity and considered only Topic, Concluding, Transition, and Supporting Sentences.

For each sample, we concatenated the paragraph P and the subsequent paragraph P', then we indicated Y by wrapping it with the special token [SENT]. We also inserted a [PARASEP] token between P and P' to indicate the paragraph boundaries. The sample was then presented as input to two language models we explored: BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019). To compute the probability of each discourse category in either model, we presented the [CLS] embedding as input to a Softmax classifier.

Model	Topic Sent.			Concluding Sent.			Transition Sent.			Supporting Sent.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	**98.85**	**97.73**	**98.29**	**11.69**	33.33	**17.31**	**7.52**	**50.00**	**13.07**	93.91	**55.10**	**69.45**
SciBERT	**98.85**	**97.73**	**98.29**	7.41	**74.07**	13.47	4.76	5.00	4.88	**94.63**	35.97	52.13

Table 4: Results on discourse category classification in terms of Precision (P), Recall (R), and F1 score.

The training objective was to minimize the following log cross-entropy loss:

$$\mathcal{L} = -\log\left(\frac{\exp(s_p)}{\sum_{j=1}^{|\mathcal{C}|}\exp(s_j)}\right),$$

where $\mathcal{C}$ represents the discourse categories of Topic, Concluding, Supporting, and Transition Sentences, s_j is the logit for the j-th discourse category label ($j = 1, \ldots, 4$), and s_p is the logit for the positive label (p is the index of the correct label).

To avoid over-representing Supporting Sentences in the Discourse Category Classification task, we balanced the label distribution in the Train and Dev sets. However, we did not perform this balancing for the Test set to determine the real-world performance of the classifiers. Note that we also tried other balancing methods, such as weighting the loss per category based on their frequency, but none worked as well.

3.2 Discourse sentence generation

To address RQ2, we investigated the influence of context on the generation of Topic (resp. Concluding, Transition) Sentences. As context we used either the remainder of the corresponding paragraph P, or we additionally included other information X, such as the Topic Ontology, or out-of-paragraph information, such as the paper's abstract and the subsequent or previous paragraph.

We describe the generation task formally here. Let P be a paragraph and let Y be a Topic (resp. Concluding, Transition) Sentence in P. The training objective is to minimize the following negative log-likelihood:

$$\begin{aligned}\mathcal{L} &= -\log p\left(Y|P \setminus Y, X\right)\\ &= -\sum_{i=1}^{|Y|}\log p\left(y_i|y_{1:i-1}, P \setminus Y, X\right),\end{aligned}$$

where y_i is the i-th token of Y, $P \setminus Y$ represents the paragraph P without Y, and X represents additional information.

We explored two classes of Pre-trained Language Models (PLMs): 1) causal language models (CLMs) that generate text in an auto-regressive manner, such as OPT (Zhang et al., 2022) and GPT-Neo (Black et al., 2022), and 2) sequence-to-sequence models (Seq2Seq) that learn mappings between the input and output sequences, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). To ensure a fair comparison, we chose models with a similar number of parameters (OPT-base and GPT-Neo both have 125M parameters, BART-base has 140M, and T5 has 220M). The inputs to all models were formed as $P \setminus Y$ concatenated with X. For CLMs, we additionally appended a separation token <|endoftext|>. For Topic Sentence generation with BART and T5, we prepended the input with "Truncated Paragraph:" and also appended "Topic Sentence:". The inputs for BART and T5 for generating Concluding and Transition Sentences were formed analogously.

For the discourse sentence generation task and each discourse category, we used only paragraphs that had at least one sentence of the corresponding discourse category, see Table 5.

Discourse category classification	Train	Dev	Test
# Topic Sentences	124	44	88
# Supporting Sentences	141	27	392
# Concluding Sentences	136	32	27
# Transition Sentences	137	31	20
Discourse sentence generation	**Train**	**Dev**	**Test**
# Topic Sentences	579	85	60
# Concluding Sentences	216	25	32
# Transition Sentences	129	34	25

Table 5: Statistics of datasets created for discourse category classification and discourse sentence generation.

3.3 Evaluation

For the discourse category classification task, we report the precision, recall, and F1 score for each discourse category. Higher scores indicate better performance. For the discourse sentence generation task, we compared the generated discourse sentences against the ground-truth sentences using summarization metrics such as ROUGE scores

(Lin, 2004) and BERTScore (Zhang et al., 2019), as well as the translation metric METEOR (Banerjee and Lavie, 2005). Higher scores indicate that the generated discourse sentences more closely resemble the ground-truths.

To quantify the readability of paragraphs, we used three automatic readability metrics, namely, Flesch-Kincaid Grade Level (FKG, Kincaid et al. (1975)), the New Dale-Chall Readability Formula (NDC, Chall and Dale (1995)), and the Automated Readability Index (ARI, Senter and Smith (1967))[4]. For these metrics, higher scores indicate higher reading difficulty and thus lower readability.

3.4 Implementation details

For the classification task, the BioBERT and SciBERT models were trained for 3 epochs with a learning rate of 2e-5, a dropout rate $p = 0.1$, and a batch size of 1.

For the generation task, all PLMs were trained for 2 epochs using the Trainer and TrainingArguments classes from the Transformers library[5]. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2e-5 and early stopping. The batch size was set to 2. For the inference step, we used beam search with $num\_beams = 3$, $top\_k = 10$, and $temperature = 0.95$.

All models were fine-tuned using a single A100 GPU provided by Google Colab. We kept batch sizes low to allow for experimenting with various context sizes.

4 Results and Discussion

4.1 PLMs accurately identify Topic Sentences

As shown in Table 4, on the discourse category classification task, both BioBERT and SciBERT achieved the highest scores of 98.29 F1 on Topic Sentences, indicating that this discourse category is the easiest to identify. Because 98.86% of Topic Sentences in our Test set were the first sentence of their respective paragraphs, a possible explanation of this finding is that the positional information of Topic Sentences can be easily captured and learned by the models.

The second-highest scores were recorded for Supporting Sentences, and the lowest scores for Transition and Concluding Sentences. We hypothesise that the poor performance on Concluding and Transition Sentences may be because both types of sentences tend to appear at the end of paragraphs, which means the model cannot rely on learning positional information alone in distinguishing the two classes. In the Appendix A, when considering Concluding and Transition Sentences as a single class, performance across all metrics improved.

Model	R-1	R-2	R-L	F_{BERT}	MTR
Topic Sentence Generation					
OPT-base	21.64	4.44	17.40	18.62	15.20
GPT-Neo	22.26	4.77	17.52	18.25	15.60
BART-base	24.33	4.72	18.49	24.67	15.39
+ PP	25.82	5.83	19.54	25.75	17.32
+ PP + A	24.90	6.15	18.98	24.78	16.88
+ PP + A + TO	**33.50**	**16.72**	**28.12**	**30.67**	**25.05**
T5-base	23.23	5.12	17.61	18.19	15.74
+ TO	30.92	15.20	26.55	24.89	23.57
Concluding Sentence Generation					
OPT-base	22.06	4.55	**18.90**	21.17	14.96
GPT-Neo	19.84	3.98	15.36	23.75	13.13
BART-base	24.11	2.84	15.55	24.52	15.39
+ PP	22.50	3.91	16.87	26.11	15.42
+ PP + A	**24.52**	**5.23**	18.84	**29.89**	**16.07**
T5-base	17.35	3.34	13.26	6.25	11.64
Transition Sentence Generation					
OPT-base	15.50	2.21	12.31	6.42	9.50
GPT-Neo	15.38	2.18	11.77	3.40	7.88
BART-base	17.00	3.27	11.35	13.99	16.33
+ NP	**23.85**	**4.51**	**15.94**	**17.80**	**19.86**
+ NP + A	21.43	3.21	14.54	13.72	15.92
T5-base	12.22	2.70	8.71	2.59	8.86

Table 6: Results on discourse sentence generation. For ROUGE scores, we report the f-measures for ROUGE-1, ROUGE-2, and ROUGE-L. For BERTScore, we report the F1 score (F_{BERT}). MTR denotes the METEOR score. PP indicates the addition of the **P**revious **P**aragraph to the input, whereas NP, A and TO indicates the addition of the **N**ext **P**aragraph, the **A**bstract, and the **T**opic **O**ntology, respectively.

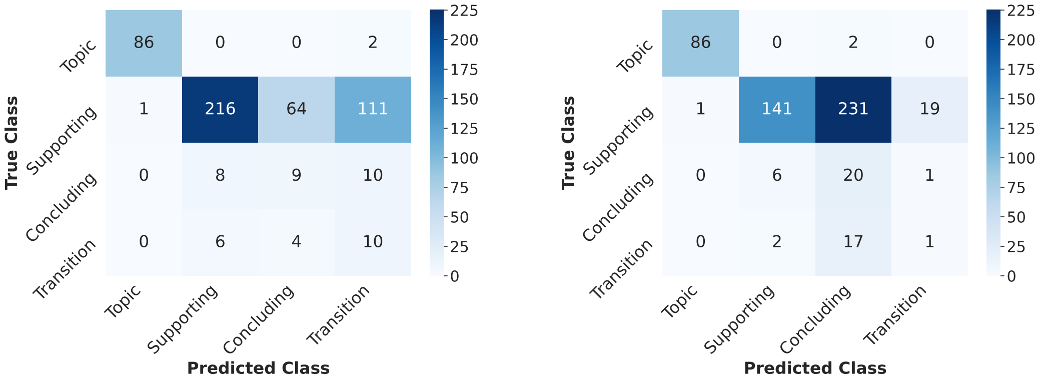
Based on the confusion matrices in Figure 3, BioBERT and SciBERT respectively tended to misclassify Supporting Sentences as Transition and Concluding Sentences. A possible explanation is that Supporting Sentences may be very diverse, and because we heavily downsampled Supporting Sentences to balance the four discourse categories for this task, our models were not able to learn this diversity.

4.2 Influence of context

On the discourse sentence generation task, both CLM and Seq2Seq models achieved the highest ROUGE F1 scores on Topic Sentences and the low-

[4] All metrics were computed with the Python package *py-readability-metrics*.

[5] https://github.com/huggingface/transformers



(a) BioBERT

(b) SciBERT

Figure 3: Confusion matrices for discourse category classification with BioBERT and SciBERT.

est scores on Transition Sentences, see Table 6. This finding was true regardless of whether context contained additional information or not, although the best generation scores across all discourse categories were achieved when additional information was included.

To delve deeper into whether sentences of a given discourse category carry information beyond the current paragraph, we conducted training of separate Seq2Seq models on text beyond the current paragraph (namely, using previous/next paragraphs and the abstract) as part of the input).

The BART model generated the best Concluding Sentences when the input contained the previous paragraph and the abstract in addition to the current paragraph. BART also generated the best Topic Sentences when the context included the Topic Ontology, abstract, and the previous paragraph.

As for Transition Sentences, incorporating the next paragraph resulted in the greatest improvement, but including the abstract deteriorated the performance. These findings suggest that pertinent information related to Topic, Concluding, and Transition Sentences can be found at diverse positions in a discourse category-dependent manner.

4.3 Trade-off between discourse structure and text readability

Text readability refers to the ease with which a reader can understand a written text (Zamanian and Heydari, 2012). The relationship between the completeness of discourse structure and text readability offers valuable insights. It sheds light on how the organization of a paragraph influences a reader's comprehension, engagement, and retention of information from a written piece.

To understand how discourse structure completeness relates to readability, we compared the readability across two groups of paragraphs: paragraphs with complete discourse structure and paragraphs with incomplete discourse structure. We filtered out paragraphs containing less than 100 words[6]. Then, we computed the readability of remaining paragraphs using the three previously mentioned metrics (FKG, NDC, and ARI).

Structure	FKG	NDC	ARI
Complete	16.75	12.68	17.98
Incomplete	***15.82**	**12.60**	***16.75**

Table 7: Readability measures for paragraphs with complete and incomplete discourse structures. Higher scores indicate that the paragraph is more challenging to read. * indicates statistical significance at $p < 0.05$.

We found that the paragraphs in SciPara are generally difficult texts to read, regardless of discourse structure completeness. This is evident by the average FKG scores of around 16 (see Table 7), which means that a university-level education would be required to comprehend these SciPara paragraphs. This result is not surprising, given that SciPara was constructed from scholarly works that are written for the scientific community.

Additionally, our results revealed that paragraphs with complete discourse structure are associated with greater reading difficulty than incompletely structured paragraphs. This is consistent with the work of Plavén-Sigray et al. (2017), who found that abstracts, which typically have complete discourse structures, are more challenging to read than the full text. As a complete discourse structure indicates a tightly connected reasoning chain, our

[6] As required by *py-readability-metrics*.

results imply a paradoxical trade-off between text readability and discourse structure: well-crafted scientific texts with complete discourse structures are inherently more difficult to comprehend.

5 Related Work

Previous works on automatic classification of discourse category of sentences from scientific papers are Dernoncourt and Lee (2017), Cohan et al. (2019), Gonçalves et al. (2020), Dayrell et al. (2012), Fisas et al. (2015), and Li et al. (2022). The discourse categories used reflected various roles within the scientific paper. For example, Dayrell et al. (2012) used BACKGROUND, GAP, PURPOSE, METHOD, RESULT, and CONCLUSION as discourse categories, and Fisas et al. (2015) used BACKGROUND, CHALLENGE, APPROACH, OUTCOME, and FUTURE WORK. Li et al. (2022) analyzed sentence roles specifically in RELATED WORK sections, introducing categories like "multi-document summarization" and "transition" for sentences bridging various topics. Our work distinguishes itself by examining discourse sentences in relation to their function in paragraph development, with annotations for "Transition Sentences" allowing us to comprehend how discourse expands over consecutive paragraphs, which is fundamental to our proposed research questions.

Moreover, there is a scarcity of research on generating sentences across these discourse categories. Shieh et al. (2019) and Song et al. (2022) conducted related studies, with the former generating abstract "conclusions" and the latter generating topic-word-constrained sentences. Our approach, however, explores generating "Topic Sentences" and other categories from the remainder of the corresponding paragraph and varying additional contexts, such as preceding paragraphs, thus addressing a research gap.

6 Conclusion

We introduced the SciPara dataset which comprises scientific paragraphs with expert annotations of sentence discourse category and of topic information. Leveraging pre-trained language models, we explored two tasks: discourse category classification and discourse sentence generation. While the models demonstrated high accuracy in identifying Topic Sentences, they encountered challenges in distinguishing Concluding, Transition, and Supporting Sentences, underscoring the inherent complexities in automating discourse category classification.

We also examined the influence of contextual input on generating discourse sentences. Our findings indicate that language models perform better with increased context, but that the context most useful depends on the sentence discourse category. For instance, Topic Ontology plays the most crucial role for Topic Sentence generation, whereas the next paragraph has the largest influence on Transition Sentence generation.

We also assessed the readability of SciPara paragraphs. Surprisingly, our analysis reveals an intriguing paradox on the relationship between discourse structure and readability. Scientific paragraphs containing at least one Topic Sentence and at least one Concluding or Transition sentence are commonly perceived as well-written. However, such paragraphs are more challenging to read.

7 Limitations

The limitations of our work include:

- SciPara is a high quality dataset. However, the acquisition of expert-annotated data is a resource-intensive process, which made expanding SciPara to a larger size difficult. This has resulted in a limited number of samples for certain discourse categories, notably Concluding Sentences (273) and Transition Sentences (188).

- Our annotation protocol exclusively targets scientific paragraphs within the INTRODUCTION and DISCUSSION sections because these sections are likely to have narrative structures. However, we refrained from including other sections due to the associated complexity.

- The readability metrics FKG, NDC, and ARI were developed to assess general domain text, not academic texts. Even so, we used them in this work because we were unable to find more fitting readability metrics.

Our future work will delve into a more comprehensive examination of the discourse structure across various sections of scientific papers. We are committed to finding innovative approaches to mitigate the cost and effort associated with human annotation, enabling the collection of a more extensive and diverse set of samples.

References

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*.

Yamen Ajjour, Johannes Kiesel, Benno Stein, and Martin Potthast. 2023. Topic ontologies for arguments. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1381–1397.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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 3615–3620.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1483–1492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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 3693–3699, Hong Kong, China. Association for Computational Linguistics.

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

Carmen Dayrell, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valéria Feltrim, Stella Tagnin, and Sandra Aluisio. 2012. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jingbing Feng, Xian Xu, and Hong Zou. 2023. Risk communication clarity and insurance demand: The case of the covid-19 pandemic. *Journal of Economic Dynamics and Control*, 146:104562.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Benjamin S. Freeling, Zoë A. Doubleday, Matthew J. Dry, Carolyn Semmler, and Sean D. Connell. 2021. Better writing in scientific publications builds reader confidence and understanding. *Frontiers in Psychology*, 12.

Yingqiang Gao, Nianlong Gu, Jessica Lam, and Richard HR Hahnloser. 2022. Do discourse indicators reflect the main arguments in scientific papers? In *Proceedings of the 9th Workshop on Argument Mining*, pages 34–50.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*, 32.

Nianlong Gu, Elliott Ash, and Richard Hahnloser. 2022. Memsum: Extractive summarization of long documents using multi-step episodic markov decision processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522.

Mateusz Hohol, Kinga Wołoszyn, and Krzysztof Cipora. 2022. No fingers, no snarc? neither the finger counting starting hand, nor its stability robustly affect the snarc effect. *Acta Psychologica*, 230:103765.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633, Doha, Qatar. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Guiming Li, Joanne Domenico, Yi Jia, Joseph Lucas, and Erwin Gelfand. 2009. Nf-κb-dependent induction of cathelicidin-related antimicrobial peptide in murine mast cells by lipopolysaccharide. *International archives of allergy and immunology*, 150:122–32.

Xiangci Li, Biswadip Mandal, and Jessica Ouyang. 2022. CORWA: A citation-oriented related work annotation dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5426–5440, Seattle, United States. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Philip M. McCarthy, Adam M. Renner, Michael G. Duncan, Nicholas D. Duran, Erin J. Lightman, and Danielle S. McNamara. 2008. Identifying topic sentencehood. *Behavior Research Methods*, 40:647–664.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

PK Ramachandran Nair, Vimala D Nair, PK Ramachandran Nair, and Vimala D Nair. 2014. Organization of a research paper: The imrad format. *Scientific writing and communication in agriculture and natural resources*, pages 13–25.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

David Nunan. 2015. *Teaching English to speakers of other languages: An introduction*. Routledge.

Shiro Otake, Shotaro Chubachi, Ho Namkoong, Kensuke Nakagawara, Hiromu Tanaka, Ho Lee, Atsuho Morita, Takahiro Fukushima, Mayuko Watase, Tatsuya Kusumoto, Katsunori Masaki, Hirofumi Kamata, Makoto Ishii, Naoki Hasegawa, Norihiro Harada, Tetsuya Ueda, Soichiro Ueda, Takashi Ishiguro, Ken Arimura, and Koichi Fukunaga. 2021. Clinical clustering with prognostic implications in japanese covid-19 patients: Report from japan covid-19 task force, a nation-wide consortium to investigate covid-19 host genetics. *SSRN Electronic Journal*.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The readability of scientific texts is decreasing over time. *Elife*, 6:e27725.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019. Towards understanding of medical randomized controlled trials by conclusion generation. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 108–117.

Tianbao Song, Jingbo Sun, Xin Liu, Jihua Song, and Weiming Peng. 2022. Topic-word-constrained sentence generation with variational autoencoder. *Pattern Recognition Letters*, 160:148–154.

Sadia Sultan and Syed Irfan. 2016. Adult primary myelodysplastic syndrome: Experience from a tertiary care center in pakistan. *Asian Pacific journal of cancer prevention: APJCP*, 17:1535–7.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

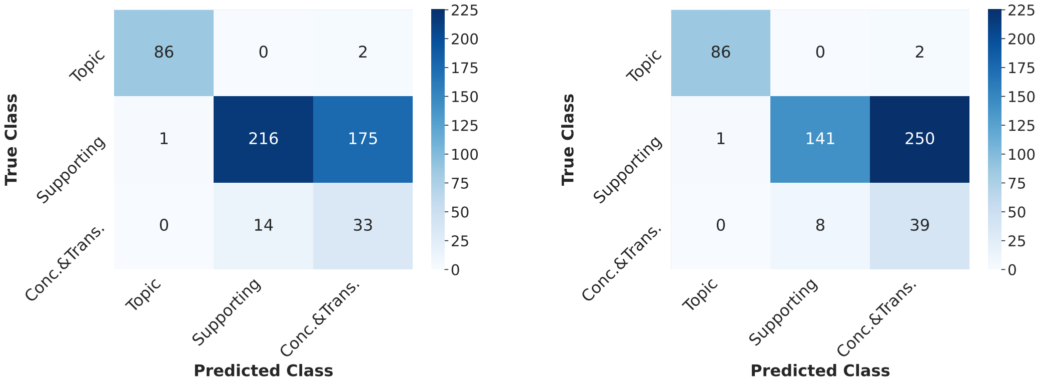
Robert C Weissberg. 1984. Given and new: Paragraph development models from scientific english. *Tesol Quarterly*, 18(3):485–500.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2:43–53.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Merged Confusion Matrix



(a) BioBERT

(b) SciBERT

Figure 4: Confusion matrix after merging the categories of Concluding and Transition Sentences.

B Dataset Example

Discourse category	Sentence
Topic Sentence	(#1) This study was the first in Japan to perform a **cluster analysis** of *COVID-19 patients*.
Supporting Sentence	(#2) We identified four clinical **sub-phenotypes**, namely the **“young healthy cluster" (Cluster 1)**, **“middle-aged cluster" (Cluster 2)**, **“middle-aged obese cluster" (Cluster 3)**, and **“elderly cluster" (Cluster 4)**, which were associated with different outcomes in Japanese *patients with COVID-19.*
Supporting Sentence	(#3) Previous reports, including ours, have shown that *comorbidities* and *mortality rates* in Japan differed from *inpatient* studies in other countries.
Supporting Sentence	(#4) Thus, the identification of the meaningful **sub-phenotypes** of *Japanese COVID-19 patients* is important.
Supporting Sentence	(#5) Notably, our study used simple baseline characteristics as variables for **cluster analysis**.
Supporting Sentence	(#6) Several previous studies have shown that **cluster analysis** is useful for **phenotyping** and predicting COVID-19 outcomes.
Supporting Sentence	(#7) However, most of these studies used complicated variables, combining a wide range of blood test results for **clustering**.
Supporting Sentence	(#8) Promptly indefinable is an important feature of defining *COVID-19* **sub-phenotypes**.
Concluding Sentence	(#9) We believe that the present simple clustering may be of great help to clinicians in predicting *prognosis* and performing individualized *therapy*.

Table 8: An example paragraph with one Topic Sentence, seven Supporting Sentences, and one Concluding Sentence. Paragraph topic is marked with bold font, while topic attributes are marked with italics. Source: Otake et al. (2021)

.

Discourse category	Sentence
Topic Sentence	(#1) Among other factors, the **SNARC effect** is considered to be **linked to the finger counting direction**.
Supporting Sentence	(#2) Fischer (2008) has shown that the **SNARC effect** was not significant (associated p-value of .061) in *participants starting finger counting with their right hand (right-starters).*
Supporting Sentence	(#3) It differed significantly from the **SNARC effect** observed in *left-starters*.
Supporting Sentence	(#4) The latter group also revealed a significant **SNARC effect**.
Supporting Sentence	(#5) Moreover, the variance in the **SNARC effect** was greater among *right-starters*.
Supporting Sentence	(#6) This observation was only partly replicated in a large-scale online study (Cipora, Soltanlou, et al., 2019), which showed a difference between *left- and right-starters* in the same direction.
Supporting Sentence	(#7) Still, it was associated with a negligibly small effect size (Cohen's d = 0.12).
Supporting Sentence	(#8) However, Bayesian analysis has shown that the result was inconclusive and was leaning towards supporting the null hypothesis.
Supporting Sentence	(#9) At the same time, unlike in Fischer (2008), a robust **SNARC effect** was found in *right-starters*, and there was no significant difference in variance between *left- and right-starters*.
Supporting Sentence	(#10) Further studies have also demonstrated a robust **SNARC** in *right-starters* (Fabbri, 2013; Prete & Tommasi, 2020).
Supporting Sentence	(#11) Additionally , in several countries where the majority of *people start finger counting with their right hand* (e.g., Belgium and Italy), the **SNARC effect** has been observed in multiple studies (e.g., Cutini, Scarpa, Scatturin, Dell'Acqua, & Zorzi, 2014; Gevers, Ratinckx, de Baene, & Fias, 2006; Mapelli, Rusconi, & Umilta, 2003).
Concluding Sentence	(#12) To sum up, there seems to be some evidence, however mixed, that finger counting is associated with the **SNARC effect** (see also Riello & Rusconi, 2011).
Supporting Sentence	(#13) Having seen these results, one might ask why the **SNARC effect** should be **related to the finger counting direction**.
Transition Sentence	(#14) The research on the *embodiment of numerical cognition* can illuminate this issue.

Table 9: An example paragraph with one Topic, Transition, Concluding, and Transition Sentence each. Paragraph topic is marked with bold font, while topic attributes are marked with italics. Links are marked with underline. Source: Hohol et al. (2022).

Discourse category	Sentence
Topic Sentence	(#1) In December 2019, a **disease outbreak** was noticed after a massive admission of *patients* with common clinical symptoms of *pneumonia* in the local hospitals of Wuhan City, China.
Supporting Sentence	(#2) Upon further investigations, the *World Health Organization* confirmed that the novel *coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)*, was responsible for these clinical symptoms and further denominated this disease as *coronavirus disease (COVID-19)*.
Supporting Sentence	(#3) Its clinical course is diverse, ranging from mild self-limited *illness* to life-threatening organ dysfunctions.

Table 10: Example badly-structured paragraph with only one Topic Sentence and two Supporting Sentences. Paragraph is marked with bold font, while topic attributes are marked with italics. Source: Otake et al. (2021).

Discourse category	Sentence
Supporting Sentence	(#1) Most published data on MDS is from Western countries.
Supporting Sentence	(#2) Published local data are scarce.
Supporting Sentence	(#3) There are few studies available from Pakistan (Irfan et al., 1998; Ehsan et al., 2010; Rashid et al., 2014).
Transition Sentence	(#4) The purpose of this study is to demonstrate demographical, clinical and the hematological features of adults primary MDS patients who visited our tertiary care center from 2010 till the end of 2014.

Table 11: An example paragraph with only one Transition Sentence and four Supporting Sentences. As this paragraph does not contain a Topic Sentence, the subsentence level part of the annotation task was not completed. Source: Sultan and Irfan (2016).

Discourse category	Sentence
Supporting Sentence	(#1) It was determined that CRAMP expression in BALB/c-derived mast cells was inducible by LPS, which also induces production of certain cytokines, including IL-13.
Supporting Sentence	(#2) This is of interest since IL-13 (and IL-14) can reportedly suppress induction of cathelicidin production by some cell types, such as antigen-exposed keratinocytes.
Supporting Sentence	(#3) In contrast, activation of mast cells with IL-4 appears to increase accumulation of cathelicidin protein.
Supporting Sentence	(#4) It was also reported that skin obtained from patients with atopic dermatitis have decreased cathelicidin LL-37 levels compared to normal skin and thus supports high levels of vaccinia virus replication, as is characteristic of eczema vaccinatum.
Supporting Sentence	(#5) Atopic dermatitis skin is characterized by overexpression of IL-4 and IL-13.
Concluding Sentence	(#6) Thus, although mast cells may be a source of cathelicidins, as described above, their presence and activation in skin could in fact, through production of certain cytokines, result in suppression of production of the antimicrobial peptides by other cell types.

Table 12: An example paragraph with only one Concluding Sentence and five Supporting Sentences. As this paragraph does not contain a Topic Sentence, the subsentence level part of the annotation task was not completed. Source: Li et al. (2009).