A corpus for Automatic Article Analysis

Elena Callegari
Uni of Iceland / Árnagarður, Reykjavík
SageWrite ehf. / Miðbær, Reykjavík
ecallegari@hi.is

Desara Xhura
SageWrite ehf. / Miðbær, Reykjavík
desara@sagewrite.com

Abstract

We describe the structure and creation of the SageWrite corpus. This is a manually annotated corpus created to support automatic language generation and automatic quality assessment of academic articles. The corpus currently contains annotations for 100 excerpts taken from various scientific articles. For each of these excerpts, the corpus contains (i) a draft version of the excerpt (ii) annotations that reflect the stylistic and linguistics merits of the excerpt, such as whether or not the text is clearly structured. The SageWrite corpus is the first corpus for the fine-tuning of text-generation algorithms that specifically addresses academic writing.

Keywords: Natural Language Generation, Automatic quality assessment of text, Scientific articles, Academic writing

1 Introduction

The latest developments in Natural Language Processing (NLP) and Natural Language Generation (NLG) demonstrate a significant gain in performance on many domain-specific NLP tasks, by pretraining on a large corpus of text and fine-tuning using prompt engineering[1] in specific task (Liu et al., 2021)(Brown et al., 2020)(Han et al., 2021). The SageWrite corpus is a manually annotated corpus created as a training dataset for the development of automatic text-generation and quality-assessment tools for academic writing[2].

When writing the different sections of an academic paper, authors often start by creating a rough draft or outline of what they want that section to say, which they then proceed to edit -and re-edit- until they are satisfied with it. An author writing the introduction of a linguistics paper may for example start by writing something along the lines of 1, which they will then proceed to edit until it looks something like 2:

1. *My intentions:*
   first: present core data on focus particles
   second, review different existing approaches
   3rd: say what I think about what works best

2. My intentions in this article are threefold: first, to outline the key data that any successful account of focus particles should explain; second, to review existing approaches that attempt to account for these data; and third, to offer my own views about the direction any successful analysis should take.

Our primary goal is automate the process that leads from 1 to 2: we want to generate grammatical text starting from a rough draft of what the final text should look like. Put differently, what we aim to do is streamline the revision process that leads from 1 to 2. What is required to generate 2 out of 1 stands halfway between natural language generation out of a limited input (Qu, 2020) and advanced automatic paraphrasing(Palivela, 2021). Our secondary goal is to develop a classifier that can process scientific articles and automatically assess whether or not they exhibit certain qualities or flaws that we deem relevant to assess scientific publications, such as whether or not information is clearly presented and whether or not the text exhibits a good flow. Again, this is in an attempt to streamline the revision process: if the stylistics shortcomings of a paper are flagged automatically, the author(s) of said paper can more readily address them. The SageWrite corpus was created to assist in the training of both of these functionalities.

[1] Prompt engineering is a way of fine-tuning, where the NLP algorithm gets fed with examples of input and expected results.

[2] In the future, the dataset could also be relevant for text summarization purposes similar to (Collins et al., 2017)

A first version of the corpus (version 0.1), consisting of 100 annotated excerpts, was published online in February 2022[3]. We plan on increasing the size of this dataset as more excerpts get annotated.

2 Text Selection

The 100 manually annotated excerpts were extracted from various types of academic articles. To obtain the excerpts, we first created a database containing scientific articles taken from Arxiv, PubMed, plus around 70 articles that we randomly selected from various disciplines in the Humanities. The articles taken from Arxiv were all dated March 2020 onwards.

To extract the excerpts, we wrote a Python program that automatically extracted excerpts of around 300 words from various points in an article. This was done to ensure that text belonging to various sections of a paper (e.g. introduction, abstract, conclusions) was included. Text was always selected from the beginning of a paragraph until the end of a paragraph. The average length of the excerpts was 193 words.

The excerpts were annotated by three annotators. As we wanted to work on academic texts, we hired annotators who had ties with academia and experience with academic writing. Accordingly, one of our annotators was a MA student, one was a university lecturer and one had a PhD degree. All annotators were also native speakers of (American) English.
Annotations were completed online on a dedicated platform where annotators could automatically log each part of the annotation for a given excerpt.

Annotators saw rotations consisting of one excerpt from a PubMed article, one from an Arxiv article and one from our Humanities articles. As we thought it would be interesting to see how different individuals would react to the same text, all annotators saw and hence annotated the same excerpts.

3 Structure of the Corpus

For each of the 100 excerpts, the corpus contains (A) three corresponding rough-draft versions of excerpt, each authored by a different annotator and (B) a list of tags that describe the stylistic and linguistic qualities of each excerpt.

[3] https://github.com/elenaSage/SageWrite0.1corpus

3.1 The Drafts

When writing up a section of an academic paper, authors generally start out by writing a rough draft of what they want to say. Drafts are both lexically and syntactically different from the final version of a paper. (Bowen and Van Waes, 2020) and (Bowen and Thomas., 2020) used the key-logging software Inputlog ((Leijten and Van Waes, 2013)) to explore how seven MA students (four native speakers of English and three native speakers of Chinese, all enrolled at a British University) approach revisions when writing an academic paper. The authors discovered that drafts feature fewer subordinates, adverbials and nominal modifiers than the finished articles. For example, some sentence-initial adverbial clauses ((underlined in 3, ex. from Bowen & Van Waes: 348) and some sentence-initial adverbials ((underlined in 4, ex. Bowen & Van Waes: 349) do not appear in the initial draft but are only added during the revision stage. Based on our own experience with academic writing, we also expect drafts to contain various types of abbreviations (e.g. 5), to be more schematic in nature (e.g. articles, copulas, 1st person singular pronouns may be dropped (6a), or arrows and empty lines may be used in place of some types of adverbials (6b)), and to contain instances of colloquial language that do not appear in the final version of a paper ((8).

3. (a) **Draft**
   "Research in this area has also looked at the differences between collectivist and individualistic countries."
   (b) **Finished Paper**
   "As well as looking at the differences in class, research in this area has also looked at the differences between collectivist and individualistic countries."

4. (a) **Draft**
   "As previously mentioned, because of the close family bond (. . .)"
   (b) **Finished Paper**
   "However, as previously mentioned, because of the close family bond (. . .)"

5. (a) Contractions: "that's" vs. "that is"
   (b) Colloquialisms: "cause" for "because", "w" for "with"
   (c) Other types of abbreviations: "mvt" for "movement", "foc" for "(linguistic) Focus"

6. (a) "in sect 1, will be talking about foc marking patterns in Malayalam"
   (b) "in sect 1 –> foc patterns in Malayalam"

7. (a) **Draft**
   "In the essay Racisms, Kwame Anthony Appiah says what he thinks about the topic"
   (b) **Finished Paper**
   "In the essay Racisms, Kwame Anthony Appiah provides his thoughts on this issue."

8. (a) **Draft**
   "But are MCI patients actually aware of their cognitive deficits? That's debatable"
   (b) **Finished Paper**
   "However, whether patients with MCI are truly aware of the full extent of their cognitive deficits is a matter of debate"

We asked our three annotators to read each excerpt and try to reverse-engineer what the draft version of that excerpt might have looked like, and to write that down. We asked them to experiment with different drafting styles; for example, we explained that while some authors might use lots of abbreviations, others might prefer to spell out every or most words. While some authors might use extremely colloquial language, others might prefer to adhere to academic lexical standards already in earlier versions of a paper.

An example of an original text and the draft one of the annotators created can be seen in 9 (original excerpt from (Keay and Hind, 2020), page 5):

9. (a) **Original Text**
   "Participants (n=88) were recruited from clients attending a private physiotherapy clinic in Bath, United Kingdom. The physiotherapy clinic provides physiotherapy, strength and conditioning programmes and clinical input for a range of conditions, including those exercisers with suspected low energy availability. Invitation for participants was also disseminated through contacts in the vicinity such as university, sport clubs and healthcare providers referring to the physiotherapy practice. The inclusion criteria were males and females over the age of 20. The study was approved by by the university research ethics committee and all participants provided informed consent prior to taking part."
   (b) **Draft Text**
   "Participants: n=88; recruited from client pool of private physio clinic, Bath, UK. Physio clinic offers physiotherapy, strength and conditioning, clinical input for many conditions, including exercisers with possible low energy availability. Also invited participants through local contacts at university, sports clubs, healthcare providers that refer to the physio clinic. Inclusion criteria: males/females >20 years Approved by uni research ethics committee; all subjects gave informed consent before participation."

As we are dealing with academic text, our goal is to develop NLG tools that do not generate too much beyond the original input: should the AI generate too much on top of the initial input provided by the user, one could question whether the resulting generated text is truly the work of the author or rather should be considered the work of the AI. Because of these concerns, we instructed our annotators not to leave out non-recoverable information from the drafts. For example, information occurring between parentheses in the original text was always included in the corresponding draft version (see 10).

10. (a) **Original Text**
    "It also presents methods that may be used for analyzing language interplays in general **(demonstrated using the PDT data)**"

    (b) **Draft Version** (as by Annotator 2)
    "Present methods to analyze language interplays in general **(see PDT)**"

Annotators first practiced annotation on a set containing 50 sample excerpts. During this practice run, annotators got direct feedback by the authors of this paper, who reviewed the annotations of the sample excerpts. These 50 practice excerpts are not included in the dataset we published online.

3.2 The Tags

We asked our annotators to evaluate the stylistic and linguistic merits of each excerpt by selecting

dedicated tags. We started out with a set of 13 tags that we came up with ourselves, based on our own personal perception of what common issues are found in scientific articles, as well as on the literature on the topic (Pinker, 2014),(Ventola and Anna Mauranen, 1996)(Badley, 2019)(Crompton, 1997). The 13 initial tags are listed below; we also provide a short explanation of those tags which may not be fully transparent.

i. Colloquial Language: to be used whenever overly colloquial language is used;

ii. Formal Language: whenever excessively formal language is used, e.g. when expressions like *et ceteris paribus* are used (too often);

iii. Jumbled Vocabulary: to describe combinations of words that make little sense, e.g. "the council has a *strong objective*"(objectives cannot be *strong*);

iv. Unnecessary jargon;

v. Verbosity;

vi. Opaque writing: for text that is obscure, hard to understand;

vii. Overly long sentences;

viii. Abuse of passive sentences: e.g. "It has been found that there had been many . . . ";

ix. Excessively complex syntax: e.g. "It is expected that an exploration of the variables affecting the effectiveness of reading aloud will support us in designing lessons (. . .)";

x. Clear Structure: to mark text that is clear and well-structured, text that clearly communicates the writer's intentions, data or results;

xi. Pretentiousness;

xii. Engaging Writing: text that is compelling, witty and makes one want to read more;

xiii. Dull writing: text that is dry, boring and not engaging;

When selecting which tags to include in our inventory, we tried including tags that refer to different linguistic dimensions. For example, tags 1 to 5 relate to the **lexical** dimension, tags 7 to 10 capture **syntactic** properties, tag 10 relates to **pragmatics** and tag 11 to 13 relate to the perceived **stylistic** merits or demerits of a text. We also tried to balance the number of positive and negative tags. We provided annotators with a document explaining each tag and where it should be used, which we went over together. We then let the annotators try out the tags over the 50 sample excerpts, providing them with personalized feedback and comments should they appear to be using some of the tags incorrectly. We also told annotators that they could suggest additional tags should they notice anything that was obviously missing. After this initial dry-run over the 50 sample excerpts, based on the suggestions from the annotators we added 6 additional tags:

xiv. Redundant (content): to be used for words, phrases or clauses that are superfluous;

xv. Repetition (style): for anaphoric repetitions, epiphoric repetitions and anytime sentence structure or vocabulary is not diverse enough; A problem which is encountered frequently in academic writing (Xiao and Carenini, 2020)

xvi. Poor flow: if the logical flow of a text is whacky, or whenever there are no clear threads to follow;

xvii. Non-sequitur: sentences that do not follow logically from anything that was said before;

xviii. Unclear/vague: for unclear referents, ambiguous statements and anything that should have been explained in more detail;

xix. Fragment: for sentences/ paragraphs that feel excessively telegraphic in style.

Also based on the suggestions from the annotators, we replaced the tag "jumbled vocabulary" with "word choice":

–> word choice: to be used for any questionable lexical choice, *whether at the sentence level or at the level of single words*.

The final tag inventory thus consisted of 19 tags. Annotators were given the option to select tags either globally or locally. Locally selected tags referred to specific sub-parts of an excerpt, e.g. to specific words, phrases, paragraphs. An example would be the tag "overly long sentences", that could apply to a single sentence. A tag that was selected globally meant that the specific characteristic that the tag singled out applied to the entire excerpt;

an example would be the tag “poor flow”. In the corpus, each excerpt is associated with each of the 19 tags, and for each excerpt each of the 19 tags has a value ranging from 0 to 3: 3 if that tag was selected for that excerpt by all three annotators, 0 if it was selected by no annotator. To simplify the structure of the corpus, we eliminated the distinction between global and local tags (in version 0.1 at least): if a tag was selected by an annotator, it is associated with a “1” value, regardless of whether the tag was selected globally or locally. The same holds for cases in which the same tag was selected locally more than once within the same excerpt. In future versions of the corpus, we plan on making the distinction between global and local tags accessible.

4 Exploratory Data Analysis

4.1 Tags used

Table 1 below illustrates how often the tags were selected at least once for a given excerpt (whether locally or globally) by an annotator. We see that the most frequently selected tags where “opaque writing” (34 instances), “clear structure” (36 instances) and “word choice” (18 instances).

Some of the tags which were relatively underused are "formal language" (1 instance), "colloquial language" (2 instances), "repetition" (2 instances) and "abuse of passive sentences" (3 instances). There are different reasons that could explain why these tags were underused: the low frequency of "colloquial language" could be explained by assuming that academic papers displaying an overly colloquial style are fairly rare; if anything, academic papers tend to be *too* formal. The low frequency of the "formal language" tag could be explained by citing difficulties in determining when text is *too* formal in a field where the use of formal language is generally encouraged. The same explanation could be extended to account for the low frequency of "abuse of passive sentences": passive sentences are a feature of academic writing. Annotators might have felt compelled to accept as good passive structures that they would have flagged otherwise precisely because they were aware they were dealing with academic text.

4.2 Length of Drafts

Figure 1 illustrates the length distribution of each of the 100 excerpts. The average length of the excerpts was 193 words.

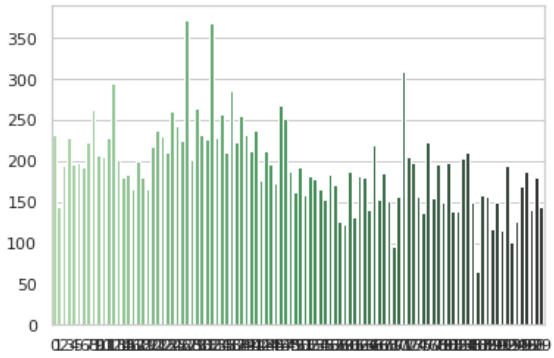

Figure 1: Length in words of each of the original 100 excerpts

Figure 2, 3 and 4 illustrate the length distribution of each of the drafts created by annotator 1, 2, and 3 respectively.

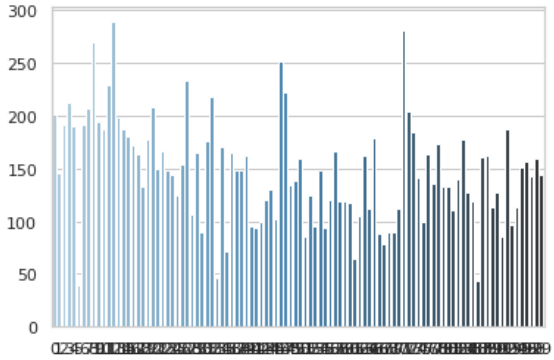

Figure 2: Length of each of the drafts created by annotator 1

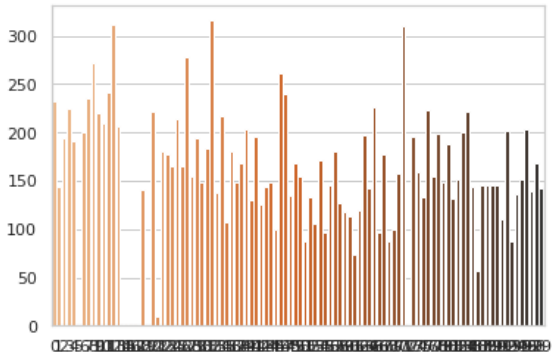

Figure 3: Length of each of the drafts created by annotator 2

Note that some of the data points are missing in figures 3 and 4 (annotators 2 and 3). This is because annotators were instructed to mark as not readable excerpts that would be too complex or time-consuming to annotate, e.g. excerpts containing lots of formulas or symbols. The missing data

colloquial language	2	abuse of passives	3	repetition	2
formal language	1	clear structure	36	fragment	9
jargon	4	pretentiousness	3	non-sequitur	4
verbosity	2	engaging	13	poor flow	8
opaque writing	34	dull	10	redundant	6
overly long sentences	4	unclear	12	complex syntax	4
word choice	18				

Table 1: Frequency of Tag Usage in corpus

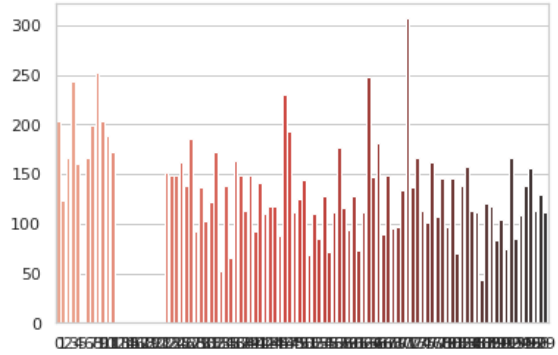

Figure 4: Length of each of the drafts created by annotator 3

points in Fig 3-4 then represent excerpts that the annotators decided to mark as not readable.

Figures 5, 6 and 7 illustrate the ratio between length of the original excerpt and the corresponding draft for each of the 3 annotators. For annotator 1, the average ratio corresponds to 0.774; for annotator 2, to 0.813; for annotator 3, to 0.633. We see that the length of a draft increases more or less incrementally with the length of the original text for annotators 1 and 2. In the case of annotator 3, on the other hand, the length of the initial outline is less reliable of an indicator of the length of the corresponding draft.

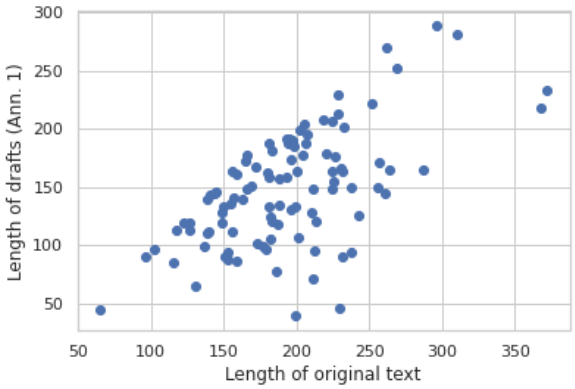

Figure 5: Ratio between length of excerpts and corresponding draft for annotator 1.

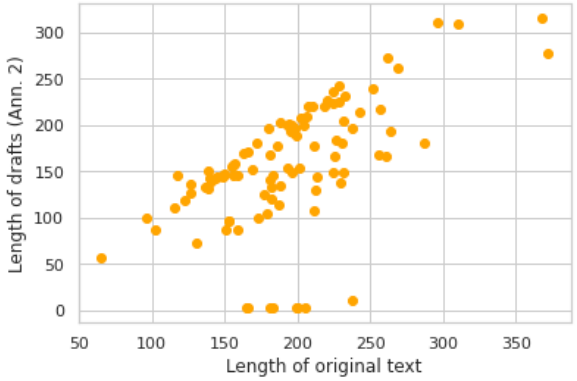

Figure 6: Ratio between length of excerpts and corresponding draft for annotator 2.

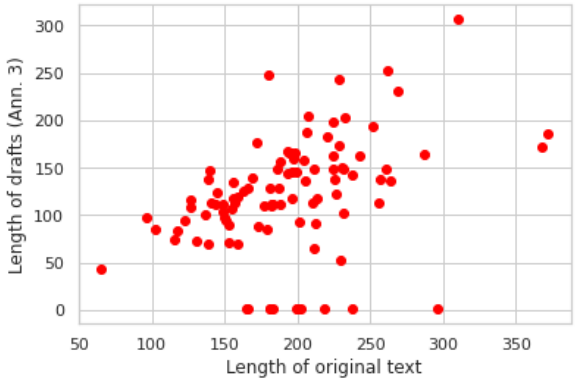

Figure 7: Ratio between length of excerpts and corresponding draft for annotator 3.

The average amount of words per draft was 146.5 for annotator 1, 155 for annotator 2 and 119 for annotator 3. Figures 8, 9 and 10 (teal scatter plots) help us further qualify these numbers by showing us how draft length compares among annotators. Figure 8 illustrates how the length of the drafts created by annotator 1 compares to those created by annotator 2. Figure 9 compares the drafts written by annotator 1 to those written by annotator 3. Finally, figure 8 compares annotator 1 with annotator 2. We see that there is indeed a difference in style between annotator 1 and 2 (some of the

drafts created by annotator 2 are longer than those created by annotator 1), and in between annotator 2 and 3 (annotator 3 writes shorter drafts). The difference between annotator 1 and annotator 3, on the other hand, seems to also be an artefact the missing data points (i.e. the original texts that the annotator decided not to annotate) for annotator 3.

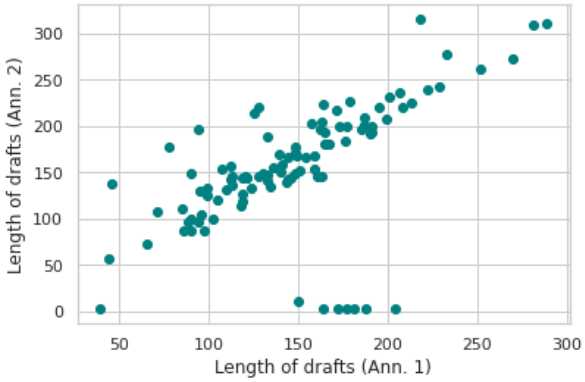

Figure 8: Ratio between length of drafts by annotator 1 and drafts by annotator 2.

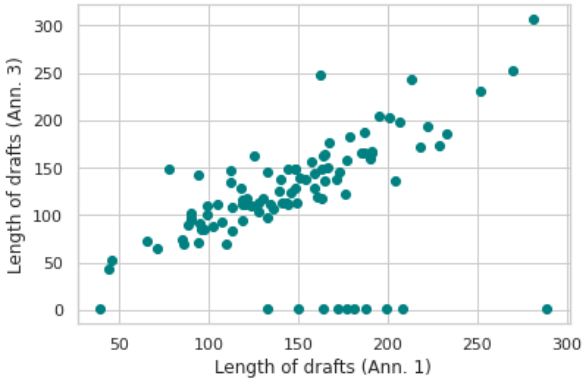

Figure 9: Ratio between length of drafts by annotator 1 and drafts by annotator 3.

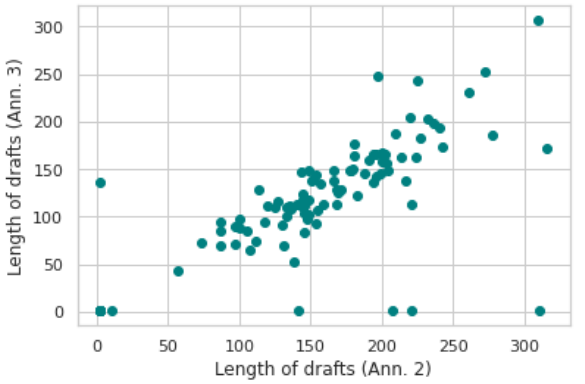

Figure 10: Ratio between length of drafts by annotator 2 and drafts by annotator 3.

We also computed the total number of words versus the number of unique words for the original excerpts, the drafts by annotator 1, those by annotator 2 and those by annotator 3 (table 2). Note that we only extracted expressions containing letter characters (symbols and digits were excluded) and with a length between 2 to 20 characters; this was mainly done to exclude formulas and mathematical symbols from the analysis. We see that the percentage of unique words over the total word count is remarkably similar overall: it is identical in both the original texts, the drafts created by annotator 1 and the drafts created by annotator 2. Annotator 3, on the other hand, appears to make a higher use of unique words. This is likely connected to the fact that annotator 3 is both a university lecturer and a writer.

5 Conclusions & Limitations

In this paper, we have illustrated the structure and creation process behind the SageWrite corpus, a manually annotated corpus created to support automatic language generation and automatic quality assessment of academic articles. Version 0.1 of the corpus contains annotations for 100 excerpts taken from various academic articles; each excerpt was annotated by three different annotators, all of whom were native English speakers. For each of these excerpts, the corpus contains (i) a draft version of the excerpt (ii) a selection of tags that reflect the stylistic and linguistics merits of the excerpt. Regarding drafts, on average drafts were around 26% shorter than the corresponding original text, although there was definitely variation among different annotators. More specifically, we saw that the ratio between length of the original excerpt and the corresponding draft was 0.77 for annotator 1, 0.81 for annotator 2 and 0.63 for annotator 3. This suggests that one should aim for drafts to be around 26±9.6% shorter than the original text; this value is particularly interesting in the context of automatically generating draft-like text from finished papers by selectively removing specific words and phrases, which is something we are also currently working on. Our data also suggests that 23.6% is a good value to aim for when it comes to lexical diversity in the drafts: this is the value obtained by computing the mean value of the lexical diversity indexes for annotators 1, 2 and 3 (22%, 22% and 27% respectively). Of interest is the fact that 22% was also the lexical diversity index of the original texts. An

	Total Words	Unique Words	% of unique words
Original texts	14187	3150	22%
Drafts by Annotator 1	12069	2753	22%
Drafts by Annotator 2	13320	2998	22%
Drafts by Annotator 3	9986	2774	27%

Table 2: Total words vs. Unique words

issue that reduces the power of our analysis is the missing data points for annotators 2 and 3: these are excerpts that the annotators decided not to annotate because they were deemed sub-optimal examples of text. This generally happened when the original excerpts contained a lot of mathematical formulas or other types of symbols. Annotators clearly disagreed on what was deemed "annotation-worthy": annotator 1 annotated all examples, while annotators 2 and 3 did not. Future rounds of annotations could be made more efficient by analyzing more in detail what kind of excerpts did not get annotated by the most stringent annotator (annotator 3 in our case), and then adjusting our extraction code to automatically exclude text that contains whatever features are common to those sub-optimal excerpts (e.g. a ratio of symbols and formulas higher than a certain value). Regarding tag usage, we saw that the most frequently selected tags where “opaque writing” (34 instances), “clear structure” (36 instances) and “word choice” (18 instances). The tags that were selected the least, on the other hand, were "formal language" (1 instance), "colloquial language" (2 instances), "repetition" (2 instances) and "abuse of passive sentences" (3 instances). To optimize future round of annotations, a possibility might be that of dropping these four tags from the list of tags annotators can choose from; this would reduce the total number of selectable tags to 15, something that would likely also simplify the annotation process.

References

Graham Francis Badley. 2019. Post-academic writing: Human writing for human readers. *Qualitative Inquiry*, 25(5):180–191.

Neil Evan Jon Anthony Bowen and Nathan Thomas. 2020. Manipulating texture and cohesion in academic writing: A keystroke logging study. *Journal of Second Language Writing*, 50:100773.

Neil Evan Jon Anthony Bowen and Luuk Van Waes. 2020. Exploring revisions in academic text: Closing the gap between process and product approaches in digital writing. *Written Communication*, 37(3):322–364.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers.

Peter Crompton. 1997. Hedging in academic writing: Some theoretical problems. *English for specific purposes*, 16(4):271–287.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future.

Gavin Francis Keay, Nicola and Karen Hind. 2020. Bone health risk assessment in a clinical setting: an evaluation of a new screening tool for active populations. *medRxiv*.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Hemant Palivela. 2021. Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1(1):100025.

Steven Pinker. 2014. Why academics stink at writing. *The chronicle of higher education*, 61(5).

et al. Qu, Yuanbin. 2020. A text generation and prediction system: pre-training on new corpora using bert and gpt-2. 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC).

Eija Ventola and eds Anna Mauranen. 1996. Academic writing: Intercultural and textual issues. 41(2).

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents.