

**Investigating Clear-Writing
Principles in Scientific Writing:
Automated with CLEARMETRICS**

B225304

8000

Master of Science

Speech and Language Processing

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2023

Abstract

This dissertation investigates the adherence to clear-writing principles in paraphrased sentences within the scientific domain. Leveraging a combination of automated systems and human surveys, the study explores the connection between writing principles, clarity, and reader preferences. A total of 740 sentence pairs were annotated with span labels and adherence flags to a set of clear-writing principles, leading to the development of the CLEARMETRICS system. This system achieved F1-scores of 74.9% and 71.7% in detecting spans (subject, verb, character, and action) in scientific and journalistic sentences, respectively, and 60.2% and 44.7% in flagging inadherence toward the specific writing rules. A human survey evaluated the effectiveness of the clear-writing rules, identifying strong preferences for certain rules, while others had minimal impact. This research revealed common challenges and offered pathways for future enhancements, contributing insights to the creation of writing tools and guidelines to enhance the clarity of scientific writing.

Acknowledgements

I would like to give my thanks to those who made this project possible:

- My supervisor, Adam Lopez, and the MSc project team, for their helpful feedback and consistency throughout the summer. I never regretted being a part of this group.
- The SLP programme, which people and events always know when to push and relax me throughout the semesters.
- The Indonesia Endowment Funds for Education, which scholarship and people have immensely helped my education and living in the UK.
- My family, who always pick up my sometimes untimely calls.
- My belief systems, that always enable me to feel grateful and hopeful despite hurdles and hard times.

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research Questions	3
1.3	Contributions	3
1.4	Roadmap	4
2	Background	5
2.1	Clarity in Scientific Communication	5
2.2	Analysing a Sentence	7
2.2.1	Dependency Parsing	8
2.2.2	Abstract Meaning Representation	8
2.3	Automatic Writing Assistance	10
2.4	Paraphrase Corpora	11
3	Augmented Paraphrase Corpora	12
3.1	Corpora Selection	12
3.2	Annotation Guideline	13
3.3	Implementation	14
3.4	Label Distribution	15
3.4.1	Span Labels	15
3.4.2	Inadherence Labels	16
3.5	Limitation	17
4	CLEARMETRICS System	18
4.1	Objectives	18
4.2	Methodology	19
4.2.1	System Design	19
4.2.2	Testing Procedures	21

4.3	Results and Interpretation	22
4.4	General Discussion	26
5	CLEARRULES Evaluation	27
5.1	Methodology	27
5.1.1	Framework	27
5.1.2	Survey Design	27
5.1.3	Sentence Selection	28
5.1.4	Data Collection	30
5.1.5	Data Analysis	31
5.2	Results and Interpretations	32
5.3	General Discussion	34
6	Conclusions	35
	Bibliography	36
A	Ten Principles for Writing Clearly	42
B	Annotation Guidelines	43
C	Research Ethics Approval	45
D	Survey Questionnaire	46
E	Data for Visualisations	56

Chapter 1

Introduction

1.1 Motivation

Effective communication is crucial in all domains, including science. Various guidebooks discuss the goals and styles of writing in scientific enterprise [Montgomery, 2017, Matthews and Matthews, 2014, Grover, 2011, Greene, 2013]. While varied in their advice, they all agree that clarity is the most important characteristic of quality scientific writing [Heard, 2014]. Clear writing enables researchers to effectively communicate complex ideas.

However, achieving such clarity in scientific writing is often faced with challenges, such as the nature of complexity in scientific discourse and the demands of precision. This often resulted in dense prose filled with jargon. Experts have prompted specific advice aimed at improving clarity, arguing that complex ideas can be expressed clearly without losing their intricacy [Gopen and Swan, 1990, Williams and Bizup, 2013].

The existing writing principles are often derived from literary or journalistic contexts. They are most likely promoted without a rigorous evaluation of their applicability in the scientific domain. To understand which principles truly enhance clarity, we need to systematically evaluate these principles in scientific contexts. Our study is designed to investigate this need, focusing on Ten Principles of Clear Writing¹, written in *Style: A Lesson in Clarity and Grace* [Williams and Bizup, 2013].

William and Bizup's principles introduce a framework for understanding and meeting reader expectations. It encourages writers to phrase the main event (action) implied in the sentence as a verb (as opposed to a noun) and assign the subject with the general doer (character). This alignment between “story” and “structure” can influence a

¹The complete list is enumerated in Appendix A.

reader’s comprehension. The following examples are taken from *Style* and we can observe the difference between a sentence that obeys this rule and the other that doesn’t. We underlined the subject, capitalised the VERB, italicised the *character*, and bold-faced the **action**.

- (1) Medieval theological **debates** often ADDRESSED issues considered trivial by modern philosophical thought.
- (2) Medieval theologians often **DEBATED** issues that modern philosophers consider trivial.

While both sentences convey the same essential idea, sentence (2) employs a more straightforward structure that clearly identifies the character “theologians” by putting the word to the subject position (instead of “debates”). It also clearly points out the characters’ actual action, which is “debated”, by putting it in the verb position (instead of “addressed”, which is more abstract).

Such revision from sentence (1) to (2) is grounded in the second and third points of the Ten Principles of Clear Writing. Other than these, we also directed our efforts towards the fifth principle, which will be covered in the next chapter of this dissertation. These three principles are enumerated into five rules, referred to as a whole in this study as **CLEARRULES**. For convenience, each rule is given a name, as shown in Figure 1.1.

CLEARRULES

2. ALIGNCHARSUBJ: Use subjects to name the characters in your story implied by the sentence;
3. ALIGNACTIONVERB: Use verbs to name their important actions;
5. Get to the main verb quickly:
 - AVOIDLONGSUBJ: Avoid long introductory phrases and clauses;
 - AVOIDLONGINTRO: Avoid long abstract subjects; and
 - AVOIDINTERRUPTSV: Avoid interrupting the subject-verb connection.

Figure 1.1: Set of writing guidance we are going to focus on.

The application of clear-writing principles might not be a straightforward task. The process requires a nuanced understanding of language and can be time-consuming. The increasing volume of scientific literature and the ever-growing importance of clear communication create a demand for more scalable solutions. Can the principles of clear writing be systemised and automated? We can divide the revision process into

two phases: (1) detecting the application of the principles in the text and (2) paraphrasing the text such that it adheres to the principles. We decided to include this inquiry in our study, focusing only on the detection phase.

We approached the task of detecting CLEARRULES adherence as also a task of detecting the subject, verb, characters, and actions of a sentence. In this study, the solution incorporates syntactic and semantic analysis of the sentence. More specifically, we utilise dependency parsing for detecting subject and verb, and abstract meaning representation (AMR) [Banarescu et al., 2013] for identifying characters and actions. While the concept and usage of dependency parse are straightforward and established, AMR is a relatively new concept, and its applications are still being explored. We studied the feasibility of these techniques for solving the detection tasks by building a system called CLEARMETRICS.

1.2 Research Questions

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

1. Does the application of CLEARRULES in scientific-themed sentences improves clarity?
2. How can we leverage natural language processing tools and algorithms to automatically check whether a sentence obeys CLEARRULES?

1.3 Contributions

This study delivers three contributions:

1. **Empirical validation of CLEARRULES.** We assess the effectiveness of CLEARRULES through a human survey. In this way, we contribute empirical evidence to validate the impact of these principles on sentence clarity and comprehension within a scientific context. This adds practical insights to a field that often relies on theoretical guidelines.
2. **CLEARMETRICS: Automation for Detecting Adherence to CLEARRULES.** To help the process of creating the human survey, we developed an automatic system for evaluating sentence adherence to the CLEARRULES. Given a sentence, the system does two things: detects spans of character, action, subject,

and verb; scores the input on adherence to CLEARRULES. This involves using natural language processing or computational linguistics concepts and tools.

3. **Augmentated Annotation for Paraphrase Corpora.** We prepare three datasets with relevant annotations for creating the CLEARRULES survey and evaluating CLEARMETRICS. There are three corpora that we annotated: ParaSCI [Dong et al., 2021]; Microsoft Research Paraphrase Corpus [Dolan and Brockett, 2005]; and newly collected examples from *Style*, our reference book.

1.4 Roadmap

The rest of this dissertation is organised as follows. Chapter 2 provides the conceptual background motivating the selection CLEARRULES and the design of the CLEARMETRICS. Chapter 3 details the process of extending paraphrase corpora with new annotations. Chapter 4 describes the methodology and results of the CLEARMETRICS system, explaining its function, design, implementation, and performance. Chapter 5 presents the methodology and results of the human survey for evaluating CLEARRULES impact. Finally, Chapter 6 draws the conclusions of this report.

Chapter 2

Background

In this chapter, we provide the necessary knowledge to motivate the methodologies of the study. We describe the importance and application of clear writing for science text, ways to analyse a sentence, and the landscape of automatic writing assistance.

2.1 Clarity in Scientific Communication

Scientific communication is vital to the dissemination of research findings and methodologies. Often obscured by dense jargon and linguistic nuances, the presentation of scientific concepts must balance word choice, conciseness, and logical flow. Clear-writing principles have been emphasised in various writing guides and manuals. Such principles enhance not only the readability of writing but also the credibility of the author [Gopen and Swan, 1990, Williams and Bizup, 2013].

Navigating the complexities of scientific communication requires a balance between conveying complex content and acknowledging readers’ cognitive processing limitations. The lack of logical flow and complex phrases in writing can make comprehension difficult. This is where the principles of clear-writing laid out in CLEAR-RULES become essential, as it puts importance in presenting the main idea early in the sentence and aligning the semantic content (“story”) and the syntactic arrangement (“structure”).

To illustrate how CLEARRULES might be applied in revising a text, pay attention to the following sentence. The subject of the sentence is distinguished by underline and the verb is written in capital.

Building upon decades of interdisciplinary research and recognizing the critical role of local engagement in sustainable change, the promotion of

democratic social change and the enhancement of community self-determination capacities EMERGE as the central objectives of community-based participatory research.

The main idea of the sentence is about the objective of community-based participatory research. To reach that main idea, a reader has to retain a lot of concepts in his memory. This happens because the sentence presents a long, abstract subject (14 words) and a lengthy introductory phrase (17 words). It leads to cognitive overload, as a study revealed that the average human's short-term memory can only hold 7 ± 2 items at a time [Miller, 1956]. The sentence structure thus challenges the reader's ability to grasp the central point efficiently. We highlight this problem in rules AVOIDLONG-SUBJ and AVOIDLONGINTRO.

Further, we might want to put “community-based participatory research” into the subject position as we can regard it as the *main character* with “(having) objective” as its *action*. To apply ALIGNCHARSUBJ and ALIGNACTIONVERB, we paraphrase the nominalised words into their verb or adjective form:

Community-based participatory research, building upon decades of interdisciplinary research and recognizing the critical role of local engagement in sustainable change, AIMS to promote democratic social change and enhance community self-determination capacities.

While the sentence is now probably easier to follow, the additional information “building upon [...] sustainable change” disrupts the connection between the subject and the verb. A clear subject-verb connection is essential for understanding the core action in a sentence. If it's disrupted, the reader may struggle to understand what is being done and who or what is doing it. Additionally, this creates awkward and jarring interruptions in the rhythm, leading to a less enjoyable and more labour-intensive reading experience. We address this issue in rule AVOIDINTERRUPTSV. Depending on the context and aims of the writer, one of the possible revisions would be using a follow-up sentence:

Community-based participatory research AIMS to promote democratic social change and enhance community self-determination capacities. The approach builds upon decades of interdisciplinary research and recognises the critical role of local engagement in sustainable change.

Above, we have demonstrated the application of CLEARRULES for revising a sentence. We can discern the general steps for revising a sentence that adheres to the rules as follows.

1. Analyse the structure of the sentence by locating the verb and the subject.
2. Discern the story or central message of the sentence, then identify the actions and corresponding characters. A character can be a concrete entity, such as “the researcher”, or an abstract concept, such as “participatory research”.
3. Align the story with the structure. This step might involve turning nominalised words into their verb or adjective form or, if necessary, introducing an explicit character.
4. Rearrange words and phrases such that the connection between concepts is more apparent, reducing the reader’s cognitive load. Reposition additional information to avoid disrupting the main subject-verb connection. This can also include breaking down a long sentence into multiple shorter ones.

The process brings more clarity without sacrificing the complexity of the original ideas; rather, it restructures them in a way that is more aligned with how readers naturally process information.

2.2 Analysing a Sentence

Our demonstration illustrates the need for writers to understand what constitutes readability. Due to the growing demand for mass communication, objective metrics have been proposed by experts. Such metrics assign numerical scores to determine the suitability of a text for a particular audience. Some common readability metrics are Flesch Reading Ease [Kincaid et al., 1975], Gunning Fog Index [Gunning, 1969], and Automated Readability Index [Senter and Smith, 1967]. Their approach is based on counting the length or the count of sentences, words, or syllables. Even though they can provide valuable insights into a text’s complexity to some extent, good readability scores don’t necessarily mean effective scientific writing. They can overlook the nuances of scientific discourse, possibly oversimplifying compound expressions or unfairly penalising complex but accurate writings. A tailored approach is needed that considers the unique demands of scientific communication.

The CLEARRULES revision steps require a deep understanding of both the syntactic and semantic aspects of a language. The syntactic analysis examines the grammatical structure and the relationships of words within a sentence. Semantic analysis,

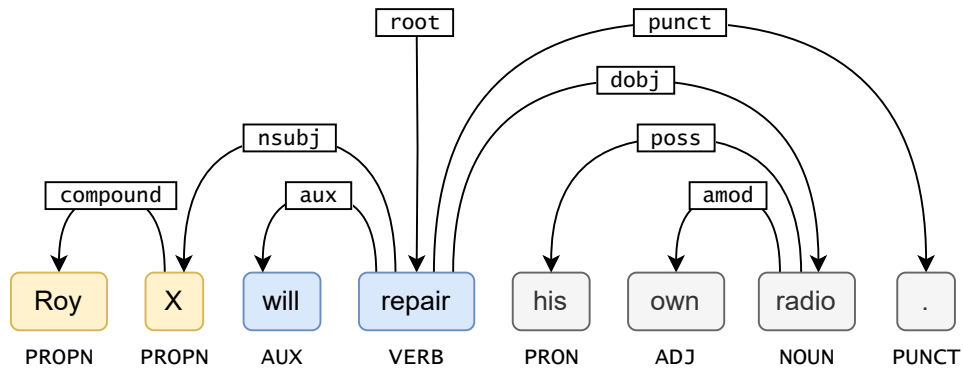


Figure 2.1: An example of dependency tree for “Roy X will repair his own radio.” along with the part-of-speech tags. The yellow-coloured and blue-coloured words are our interpretation of the subject and the verb, respectively.

on the other hand, aims to understand the underlying meanings, intentions, and relationships between concepts embedded in a text. By employing these analyses, an automated system could theoretically replicate the manual task of detecting subjects, verbs, characters, and actions of sentences.

2.2.1 Dependency Parsing

One of the kind of syntactic analysis is dependency grammar, which is usually illustrated as a tree. Dependency is the notion that words or other linguistic units relate to each other through directed links. In dependency grammar, except the root of a sentence, each word has a *head*, and each word may also function as the head for other words, called *dependents*. Components of a dependency tree are nodes of the individual words and edges between them, which are labelled with the grammatical relationships, or *dependencies* [Jurafsky and Martin, 2023]. One of the most popular frameworks for dependency grammar is Universal Dependency (UD) [Nivre et al., 2017]. An example sentence described using UD notations is shown in Figure 2.1.

In the example, we see that “repair” has no head, making it the root of the sentence. The root word usually is the sentence’s predicate or main verb. As for the subject “Roy X”, it is connected to the root by *nsubj* relation.

2.2.2 Abstract Meaning Representation

Unlike dependency relation, abstract meaning representation (AMR) is a semantic representation language that captures the meaning of a sentence in a graph-based structure

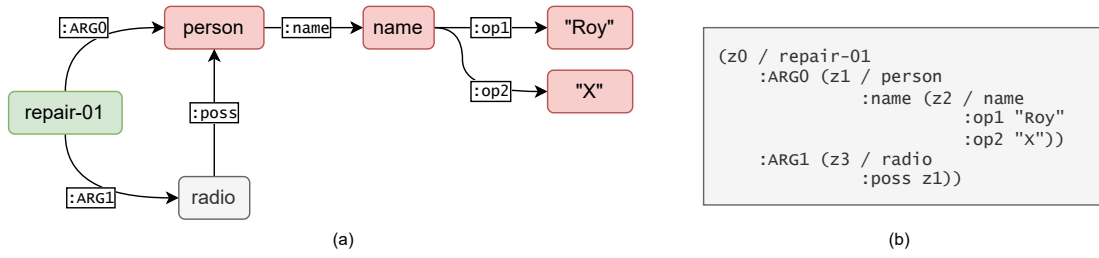


Figure 2.2: An example of abstract meaning representation graph (a) and its linearised form (b) for “Roy X will repair his own radio.” The red-coloured and green-coloured nodes are our proxy for the action and the character of the sentence.

[Banarescu et al., 2013]. In AMR, the nodes represent the concepts, while the directed, labelled edges illustrate the relationships between these concepts. The structure of AMR intends to be language-independent, focusing on the core meaning rather than specific linguistic constructs. By stripping away grammatical information, AMR offers a simplified and more universal understanding of a text, making it useful for tasks such as text summarisation [Liao et al., 2018] and paraphrasing detection and generation [Issa et al., 2018, Huang et al., 2023]. However, the abstraction level in AMR also means that certain syntactic details might be lost, such as tense and word order. Figure 2.2 illustrates how the future tense “will” is an essential part of the meaning, expressing a future intention or plan. If we back-translate the AMR graph, “Roy X repaired his own radio” is also regarded as a correct interpretation. This means AMR could lead to a less nuanced understanding of the original sentence’s meaning.

AMR convention was influenced by PropBank, a resource that provides verb-specific semantic role labelling [Palmer et al., 2005]. In PropBank, propositions or predicates are expressed in their verb form added with numbers. Each predicate has arguments or roles, typically labelled as ARG0, ARG1, etc. These arguments describe the relationships between the verb and other elements of the sentence. In Figure 2.2(b), the verb “repair” is represented by the concept `repair-01`. The ARG0 role corresponds to the agent of the action, in this case, the `person` concept named “Roy X”. The ARG1 role corresponds to the theme or object of the action, in this case, “his own radio”, which is presented by the `radio` concept with a possessive relation to “Roy X”.

Aligning AMR to the equivalent sentence is an essential step in working with AMR. The alignment connects the elements of an AMR graph to the words or phrases in the corresponding word or token in the natural language text. The alignment can be challenging due to the abstraction process and the non-linear structure of the graph.

Automated methods and tools have been developed to facilitate this alignment [Liu et al., 2018, Martínez Lorenzo et al., 2023]. These methods usually are language family-specific and might not work well with other language families with different characteristics [Anchiêta and Pardo, 2020, Oral and Eryiğit, 2022].

2.3 Automatic Writing Assistance

Following the `CLEARRULES` revision steps is a great practice for writers who want to seriously improve their style. But since scientific progress is being promoted globally [Chu and Evans, 2021], automating writing assistance is on demand. Solutions have been brought by the recent advancements in natural language processing (NLP), creating a niche field called automatic writing assistance.

Writing is a complicated process that could be assisted in many different aspects using NLP tools. Some efforts to engage researchers in this problem are the creation of shared tasks for revising a text [Dale and Kilgarriff, 2011], correcting grammatical errors [Ng et al., 2013], and predicting the intention of writers when revising a text [Daudaravičius, 2015]. NLP researchers have framed automatic writing assistance into various tasks: modelling revision steps [Du et al., 2022, Kim et al., 2022] which is used in Grammarly¹; utilising large language model (LLM) for suggesting the next words for language learners [Gayed et al., 2022] as well as academic writers [Ito et al., 2020]; and prompting LLMs to revise a text with specific edit instructions [Dwivedi-Yu et al., 2022].

From this short examination of the topic, two main schools of text revision have emerged. At one end of the spectrum, there are tools focusing on general aspects of writing such as grammatical corrections, spell checks, and span-level edits. These tools often provide broad assistance but may overlook specific nuances and stylistic considerations that are vital in scientific writing. On the other end, highly specialised tools employ language-model-based methods for next-word suggestions or specific textual generations. While these offer more tailored guidance, they might become overly prescriptive, constraining the writer’s voice and possibly hindering the creative process. The approach we propose in this study falls between these two extremes, offering guidance that is both targeted and flexible. By identifying violations of specific writing principles without dictating exact solutions, it invites writers to engage in a reflective

¹<https://www.grammarly.com>

revision process that respects individual expression while adhering to recognised standards of clarity.

2.4 Paraphrase Corpora

As we explore the various tools and techniques employed to enhance writing quality, it's crucial not to overlook the fundamental task of paraphrasing. Paraphrasing is a task that involves rewriting a text so that it conveys the same meaning using different words or structures. In the context of natural language processing, the automation of paraphrasing tasks can also be applied to text summarisation, machine translation, and question-answering.

Several corpora have been developed to facilitate research in paraphrasing. To name a few, there are Microsoft Research Paraphrase Corpus [Dolan and Brockett, 2005], which presents pairs of sentences from news articles that can be identified as paraphrasing or not; Quora Question Pairs², a collection of similar question posted in Quora; ParaNMT [Wieting and Gimpel, 2018], consists of pairs of sentences that generated through back-translating non-English sentences; and ParaSCI [Dong et al., 2021], includes paraphrased sentences from science papers. Computational linguists have added additional annotations on top of paraphrase corpus to incorporate more meaningful information that can be useful for other downstream tasks such as plagiarism detection [Vila et al., 2014, Kovatchev et al., 2018, Alvi et al., 2021]

²<https://data.quora.com/First-Quora-Dataset-Release-QuestionPairs>

Chapter 3

Augmented Paraphrase Corpora

Previously, we have established the need for CLEARMETRICS system to automate the selection of sentences that exhibit the application of CLEARRULES or lack thereof. We choose paraphrase corpora based on the design of the survey, which will be explained in Chapter 5. Three different sources were augmented with new annotations, resulting in three datasets that will become the basis of CLEARMETRICS development.

3.1 Corpora Selection

Inside a paraphrase corpus, usually, there are sets of at least two sentences that are deemed as a paraphrase of each other.

The three corpora we annotated are as follows.

1. **StyleExamples**: 70 pairs of sentences which we collected from the examples and exercises in *Style: A Lesson in Clarity and Grace* [Williams and Bizup, 2013]. We put the sentences that are deemed to need revising by the author of the book into the `inadherent` subset, while the exemplary sentences that adhere to the CLEARRULES were put into the `adherent` subset. The content of the texts is mainly about writing, business, and healthcare.
2. **ParaSCI**¹ [Dong et al., 2021]: 200 pairs of paraphrasing sentences, with 150 pairs are a subset of ParaSCI-arXiv train set and the rest 50 pairs are a part of ParaSCI-arXiv test set. The sources of the texts are ArXiv Bulk Data² and S2ORC [Lo et al., 2020]. The main topics are physics, computer science, and

¹<https://github.com/dqxiu/ParaSCI/tree/master/Data>

²https://info.arxiv.org/help/bulk_data/index.html

engineering. We do not modify the `source` and `target` sets of the pairs in the original corpus.

3. **MSRP** (Microsoft Research Paraphrase Corpus)³ [Dolan and Brockett, 2005]: 150 pairs of sentences. Since the original corpus features both paraphrase and non-paraphrase pairs, we only select the first 100 paraphrase pairs of the train set and the first 50 paraphrase pairs of the test set, respectively. The sentences were from articles on the web. The topics are mostly politics, finance, and general news. Also, we do not modify `sentence 1` and `sentence 2` of the pairs in the original corpus.

Although the development of CLEARMETRICS does not require any model training (as detailed in Chapter 4), we still prepared a train set to adhere to the conventional structure of a machine learning dataset. In this study, the train set is not used for its traditional purpose but rather serves as data for system validation. The test sets are employed to report the evaluation metrics. The approach maintains a familiar framework while adapting to the unique needs of our analysis.

3.2 Annotation Guideline

Each corpus is augmented with new annotations that mark the spans of the character, action, subject, and verb of each sentence. As an illustration, see the example in Figure 3.1. The subjects and verbs usually span into multiple words, consisting of the whole phrase that includes them. In contrast, characters and actions labels are only given into the “main” words and omit grammatical details.

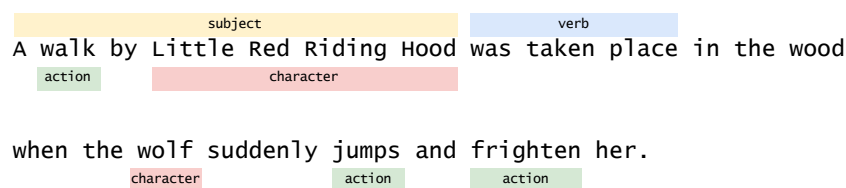


Figure 3.1: The desired span labelling.

We ask the annotator to add relations between the subjects and characters to their corresponding verbs and actions, respectively. While it’s not necessary to do that in

³<https://www.microsoft.com/en-us/download/details.aspx?id=52398>

the scope of this project, the extra information is helpful for quality-checking the annotation work.

Additionally, the sentences are also marked whether they are adhering to the CLEAR-RULES, which is called the inadherence label. This is framed as multi-classification labels, which labels are structured according to the individual points in CLEARRULES, but rather than marked compliant sentences, the annotator marked the violations towards the rules. In the example sentence above, the annotator has to mark the sentence as violating `ALIGNACTIONVERB` because the main action “walk” is not within the verb “was taken place.”

For the complete annotation guideline, see Appendix B.

3.3 Implementation

The annotation process was conducted using Label Studio⁴, an open-source data labelling platform. Each sentence was annotated individually, with the interface depicted in Figure 3.2.

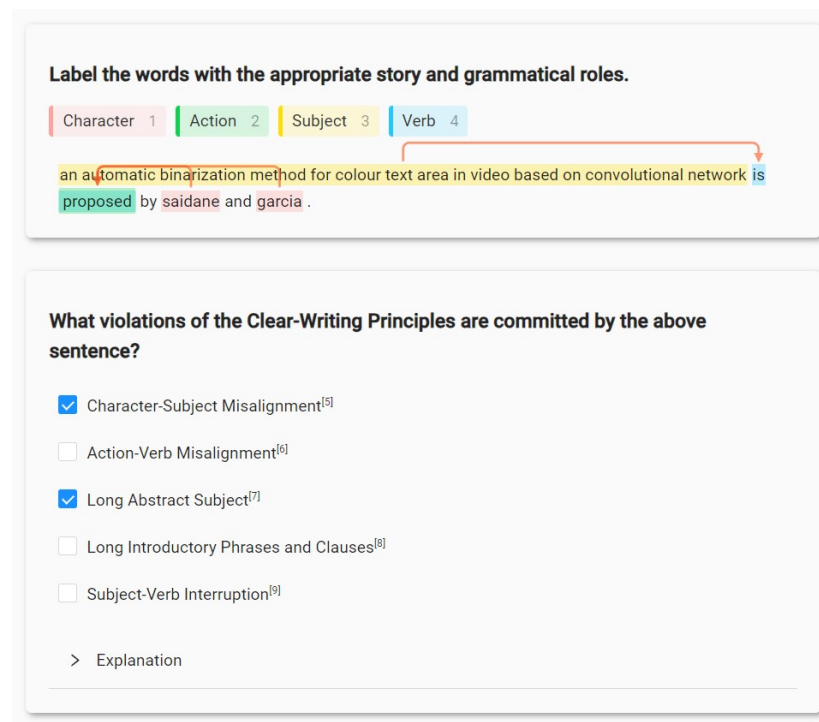


Figure 3.2: Label Studio interface for annotating the sentences.

The annotator is considered an expert in the problem, that is the author of this

⁴<https://labelstud.io/>

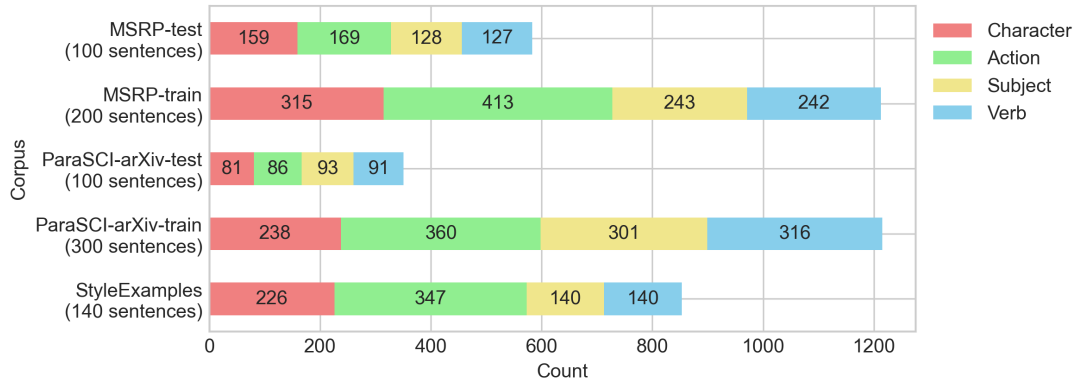


Figure 3.3: The frequency of span labels across datasets.

study. The annotator adhered to the initial annotation guideline, accessible from the interface. However, as edge cases were identified throughout the process, we consulted and further refined the guideline with reference to the book *Style: A Lessons in Clarity and Grace* [Williams and Bizup, 2013].

It’s important to note that the annotations were completed by a single individual, without an extensive evaluation of annotation quality other than several sweeps of re-checks. This aspect represents a limitation of the study and offers an avenue for future improvement.

3.4 Label Distribution

Assessing the distribution of labels within the augmented paraphrase corpora is essential for understanding the balance and variety of annotated data, which can influence the performance of models trained on this corpus.

3.4.1 Span Labels

The distributions of span labels, illustrated in Figure 3.3, reveal a consistency between the training and test sets for both corpora. This suggests that the test sets are representative of the train set, a crucial aspect for valid model evaluation.

We found interesting patterns from a comparative analysis between the ParaSCI-arXiv corpus, sourced from scientific papers, and the MSRP corpus, derived from web news. ‘Action’ words are more frequent in MSRP (1.94 per sentence) than in ParaSCI-arXiv (1.12 per sentence). A similar pattern is found with ‘Character’ words, reflecting the differing communicative purposes in scientific and journalistic writing. While

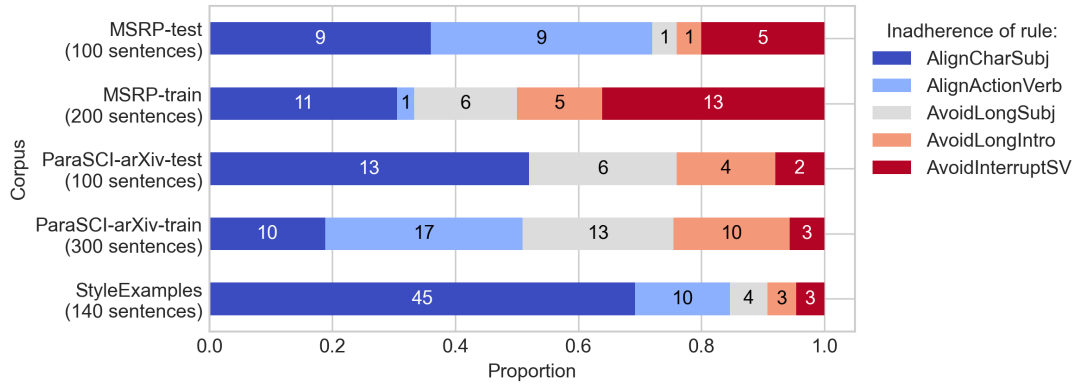


Figure 3.4: The distributions of the inadherence labels across datasets.

scientific texts prioritise precision and complexity, we news utilises more vivid and action-oriented language.

A notable characteristic of the MSRP corpus is the exceedance of subjects and verbs over the total number of sentences. The relative frequencies of subjects and verbs per sentence are 1.24 and 1.27, respectively. This pattern is attributed to the practice of quoting individuals, where both the person quoted and the subject of the quote are annotated as subjects. For example, in the sentence “*This deal makes sense for both companies,*” *Halla said in a prepared statement.*, both “This deal” and “Halla” would be annotated as subjects.

3.4.2 Inadherence Labels

The distribution of the CLEARRULES-inadherent labels is depicted in Figure 3.4. We will explain a few insights for each of the labels.

For ALIGNCHARSUBJ inadherence, it appears to be more frequent in the StyleExamples set compared to the other datasets. StyleExamples, being a collection to illustrate specific writing principles, provides significantly more examples for the most fundamental principle in the Ten Principles of Writing Clearly from the book *Style*.

For ALIGNACTIONVERB inadherence, the misalignment occurs more in the ParaSCI-arXiv train set than in both MSRP datasets. Interestingly, in the ParaSCI-arXiv test set, the label does not occur at all. This inconsistency might affect the reliability of the datasets to accommodate the learning and evaluation of a model.

For AVOIDLONGSUBJ and AVOIDLONGINTRO inadherence, the labels are slightly more frequent in the ParaSCI-arXiv sets, possibly reflecting the formal and complex sentence construction often found in scientific literature. In contrast, the AVOIDLONG-

INTRO inadherence labels appear more in the MSRP datasets. Journalistic writing might often use additional information or clauses between the subject and verb to provide context or emphasize a point.

3.5 Limitation

In the context of evaluating the CLEARMETRICS system, the augmented paraphrase corpora were carefully examined, revealing a noticeable class imbalance. Even though this could be a mirror of real-world writing practice, the unequal representation of CLEARRULES-adherent and inadherent sentences might hold significant implications when evaluating the CLEARMETRICS system. The under-representation of various inadherence classes may limit the generalizability of the findings. The sentences selected might not be varied enough to represent all cases, leading to potential biases or misinterpretations.

The recognition of this limitation can guide future refinements of the study, including considerations for more balanced data that better captures the varied application of clear writing principles for scientific writing.

It is also recommended to invite more annotators to further improve data reliability. When there are at least two annotators working on the same sentences, an objective evaluation such as inter-annotator agreement [Artstein, 2017] could be employed to measure the quality of the annotations.

Chapter 4

CLEARMETRICS System

Although the study initially utilised manually annotated corpora to identify adherence or violations, the process was found to be time-consuming. CLEARMETRICS was introduced to quickly scan sentences, generating a binary vector called CLEARFLAGS. The system’s role in this study was to filter sentences in the corpora to prepare for the CLEARRULES evaluation survey. While the manual annotations provided accurate data, CLEARMETRICS’s automation added efficiency, assisting in the preparation of the survey materials.

4.1 Objectives

To validate the CLEARMETRICS system’s efficacy and explore its potential, our evaluation process revolves on three primary objectives:

1. **Understanding AMR’s role in character and action detection:** to validate whether Abstract Meaning Representation (AMR) can effectively identify characters and actions in inadherent sentences as well as in adherent sentences. This will confirm or refute our assumption that AMR is capable for the task.
2. **General span labelling performance:** to assess how well CLEARMETRICS tags character, action, subject, and verb spans across different corpora. Understanding the system’s performance in tagging these labels will provide insights into its robustness and potential real-world application.
3. **Classification of sentence adherence to CLEARRULES:** to gauge the system’s ability to classify sentences based on their adherence to CLEARRULES. This

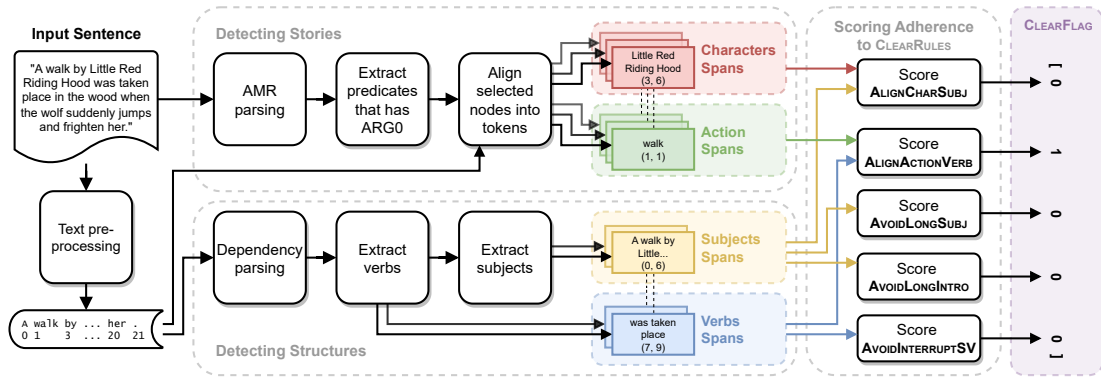


Figure 4.1: The overview of subsystems and their pipeline within CLEARMETRICS.

evaluation will determine how accurately CLEARMETRICS can identify violations of the writing principles, which is central to its intended use.

4.2 Methodology

4.2.1 System Design

As shown in Figure 4.1, the system design for CLEARMETRICS can be divided into three main parts:

1. Detecting “Structures”

- **Goal:** Identify the spans of subjects and verbs within the sentence.
- **Method:** The main verb is obtained from the root of the sentence’s dependency tree. Any conjugate verb phrases are taken as additional verbs. The noun phrases inside each subtree of the verbs are identified as subjects.
- **Output:** Spans are stored in terms of token numbers, marking the start and end of the identified subjects and verbs.
- **Implementation:** We utilise the preprocessing tools and the dependency parser from SpaCy¹, specifically using the `en_core_web_md` model [Honnibal and Johnson, 2015]. The model is chosen because, from our initial tests compared to other models and frameworks, it has the balance between accuracy, inference time performance, and easy implementation.

2. Detecting “Stories”

¹<https://spacy.io/>

- **Goal:** Identify the spans of the characters and actions within the sentences.
- **Method:** AMR parser is utilised to find propositions/predicates that have ARG0 relations. In the AMR annotation convention, predicates are used to describe events or states of affairs, so this is for representing action. The ARG0 relation generally refers to agents, causers, or experiencers of the events, thus, making the related concepts the natural candidates for characters. The selected AMR nodes then aligned the sentence, connecting the characters and actions to the tokens.
- **Output:** The start and end of each span are stored, indicating the positions of characters and actions.
- **Implementation:** For parsing the text to AMR, we utilise SPRING (Symmetric ParsIng aNd Generation) [Bevilacqua et al., 2021], which used to be a state-of-the-art model, achieving 83.0 Smatch. For the alignment task, we use a rule-based word aligner based on JAMR aligner [Liu et al., 2018]. The model choices are pragmatic since they are easily accessible through AMRLib package².

3. Flagging Inadherence to CLEARRULES

- **Goal:** Assess the sentence’s violation of the CLEARRULES based on the spans detected in the previous steps.
- **Method:**
 - (a) ALIGNCHARSUBJ: The system checks whether any character span is a subset of any subject span. If found, return 0; otherwise, return 1.
 - (b) ALIGNACTIONVERB: Check whether any action span is a subset of any verb span. If it is found, return 0; otherwise, return 1.
 - (c) AVOIDLONGSUBJ: Ensure that the first subject does not exceed a certain length. Count the number of tokens in the first subject span. Return 0 if the count is less or equal to the threshold, that is 8; otherwise, return 1.
 - (d) AVOIDLONGINTRO: Ensure that the number of tokens before the first subject span does not exceed a certain length. Count the number of tokens before the first subject span. Return 0 if the count is less or equal to the threshold, that is 8; otherwise, return 1.

²<https://amrlib.readthedocs.io/en/latest/>

- (e) **AVOIDINTERRUPTSV**: Ensure that interruption between subjects and verbs is limited. Count the total tokens of interruption between the subject and verb for all subject-verb pairs. Return 0 within the tolerance, that is 4; otherwise, return 1.

If no story or structure element is found in the sentence, then the affected rules return 0. The thresholds and the tolerance are approximations that we think made sense, based on the 7 ± 2 findings [Miller, 1956]. The counting excludes punctuation and treats the span of an entity as only one count. This decision reflects how readers generally process entity: as one cohesive idea rather than separate words. Meanwhile, punctuation marks, such as commas, are structural components and don't generally contribute to the cognitive complexity of understanding a chunk of text; instead, they aid comprehension and generally should not be penalised.

- **Output**: a binary vector called **CLEARFLAGS**, where each element corresponds to a specific **CLEARRULES**, with 0 indicating adherence and 1 meaning inadherence.
- **Implementation**: Using the spans from the previous steps, the checks are being done using Python's loops and if-else statements.

The design of CLEARMETRICS embodies an explainable approach that seeks to disentangle the syntax and semantics of a sentence. This approach allows for a nuanced understanding of sentence construction and offers a transparent mechanism for evaluating adherence to CLEARRULES.

4.2.2 Testing Procedures

Below are the testing procedures to answer the outlined objectives, including the reason for selecting the evaluation metrics.

1. Understanding AMR's role in character and action detection.

- **Test sets**: StyleExamples `inadherent` and `adherent` set.
- **Metrics**: Precision, recall, and F1-score evaluated on span-level exact match. These metrics are commonly used to evaluate classification tasks, providing a balanced view of the model's performance. Precision measures the

accuracy of positive predictions, recall assesses how well the model identifies all relevant instances, and the F1-score combines these to provide a single performance figure. A confusion matrix is further advised to check the system performance on the inadherence flagging task.

2. General span labelling performance.

- **Test sets:** ParaSCI-arXiv test set and MSRP test set, annotated for character, action, subject, and verb spans.
- **Metrics:** Precision, recall, and F1-score.

3. Classification of sentence adherence to CLEARRULES.

- **Test sets:** ParaSCI-arXiv test set and MSRP test set, annotated for adherence to CLEARRULES.
- **Metrics:** Precision, recall, and F1-score.

4.3 Results and Interpretation

Understanding AMR's role in character and action detection

The performance in subject detection is notably strong in both the `adherent` set (84.2% F1-score) and `inadherent` set (87.3%), as shown in Table 4.1. As for verb detection, we have fairly consistent results in `inadherent` set (83.8% F1-score), while the `adherent` has a lower score (76.2%) due to its lower precision. In contrast, the overall scores for detecting character and action are lower, with the result in `adherent` sentences (66% and 64.7% F1-score, respectively) being better compared to `inadherent` ones (53.2% and 53.3%, respectively). While this indicates an ability to recognise constructs that align with clear writing principles, it also raises questions about the system's effectiveness in flagging non-adherent text, the primary function of CLEARMETRICS.

We can confirm that inquiry by observing the confusion matrices in Figure 4.2.

Interestingly, we have an observation that seems somewhat contradictory. The low character span recall in `inadherent` sentences (46.7%) would imply that the system wouldn't be able to find the intersection between character span and subject span because the character span itself is nonexistent. Yet we have high false negatives (21) for `ALIGNCHARSUBJ` inadherence, meaning that the system often fails to recognise

	Precision		Recall		F1-score		Support	
	a	b	a	b	a	b	a	b
Character	61.8	72.1	46.7	60.9	53.2	66.0	135	174
Action	62.7	68.7	46.4	61.2	53.3	64.7	138	165
Subject	82.1	80.0	86.5	96.0	84.2	87.3	74	75
Verb	79.5	67.8	88.6	87.1	83.8	76.2	70	70
Micro Avg	71.4	71.7	60.7	70.2	65.1	71.0		

Table 4.1: CLEARMETRICS performance on StyleExample *inadherent* (a) and *adherent* (b) sentences

violations even when they are present. This phenomenon might be explained by the flaw in the method of the system’s flagging, specifically when there’s no story or structure element found in the sentence. If either character or subject span is not found, the system will always flag the sentence as *adherent* (negative). This overly general exception rule might contribute to the high false positives.

For the *adherent* set, the system often wrongly flags the sentences as being *inadherent* to ALIGNCHARSUBJ and ALIGNACTIONVERB (22 and 19 false positives, respectively). The high false positive rate might indicate an over-sensitivity in the system’s criteria for flagging violations. So far, we cannot pinpoint the most possible explanation, as the performance of the character and action detection themselves are not too impressive.

As for subject or verb-related rules, we observe much fewer false positives and false negatives across the sets.

Generally, the CLEARMETRICS system appears more aligned with CLEARRULES-*adherent* sentences, showing a tendency to better detect and label spans within these sentences. In contrast, the system’s capability to identify *inadherent* features and flag violations is more nuanced and shows room for improvement. This observation may underline the importance of refining the system’s approach to identifying non-*adherent* structures, potentially through fine-tuning the thresholds or other parameters.

General span labelling performance

Based on Table 4.2, the following highlights the performance of CLEARMETRICS on two different datasets, ParaSCI-arXiv and MSRP test sets, across different span labels.

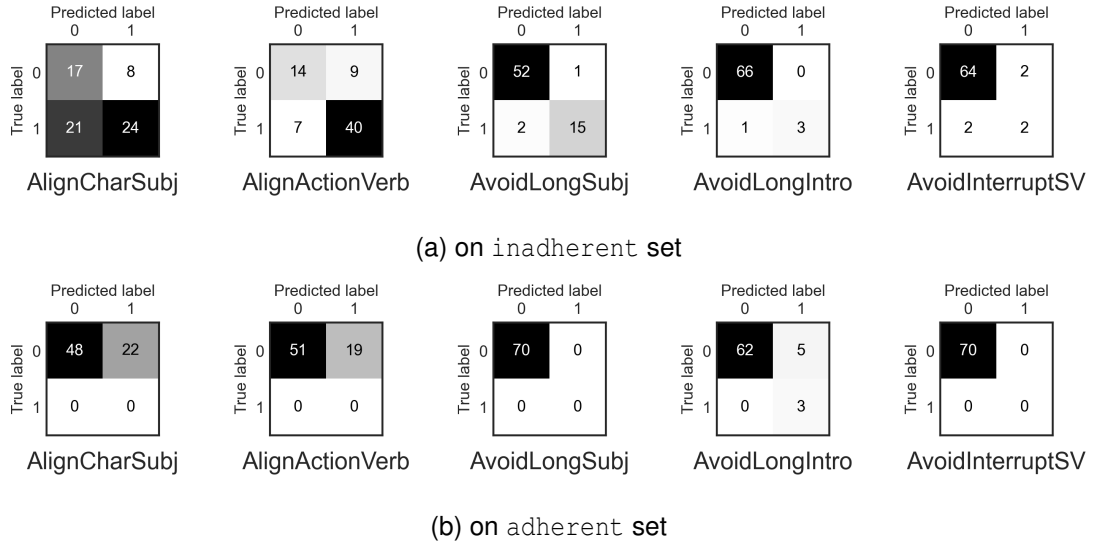


Figure 4.2: Confusion matrix of CLEARFLAGS evaluated on each StyleExamples set.

Corpus	Span Label	Precision	Recall	F1-score	Support
ParaSCI-arXiv	Character	62.0	81.6	70.5	98
	Action	56.6	80.2	66.4	91
	Subject	89.8	86.3	88.0	102
	Verb	73.5	79.1	76.2	91
	Micro Average	68.9	81.9	74.9	
MSRP	Character	57.9	56.2	57.1	176
	Action	80.1	66.8	72.9	205
	Subject	81.1	87.6	84.2	137
	Verb	70.3	81.9	75.6	127
	Micro Average	72.1	71.3	71.7	

Table 4.2: The results for span labelling across different corpora (ParaSCI-arXiv and MSRP) and four span categories

CLEARMETRICS in the ParaSCI-arXiv dataset highlights high recall across character and action spans, at 81.6% and 80.2%, respectively. Unfortunately, the precision for both spans dropped by around 20 points compared to their recall scores. On the other hand, the highest and most robust performance is seen in the subject and verb span with an F1-score of 88.0% and 76.2%, respectively.

Corpus	Inadherence to	Precision	Recall	F1-score	Support
ParaSCI-arXiv	ALIGNCHARSUBJ	64.3	69.2	66.7	13
	ALIGNACTIONVERB	40.0	60.0	48.0	10
	AVOIDLONGSUBJ	85.7	85.7	85.7	7
	AVOIDLONGINTRO	75.0	75.0	75.0	4
	AVOIDINTERRUPTSV	14.3	50.0	22.2	2
Micro Average		53.2	69.4	60.2	
MSRP	ALIGNCHARSUBJ	23.1	33.3	27.3	9
	ALIGNACTIONVERB	60.0	70.6	64.9	17
	AVOIDLONGSUBJ	33.3	50.0	40.0	2
	AVOIDLONGINTRO	27.3	75.0	40.0	4
	AVOIDINTERRUPTSV	20.0	40.0	26.7	5
Micro Average		36.8	56.8	44.7	

Table 4.3: The results for inadherence detection across different corpora (ParaSCI-arXiv and MSRP) and five inadherence categories.

In the MSRP dataset, the results are a little bit different. The precision of action detection is higher than their recall (80.1% vs. 66.7%). Character detection also follows a similar pattern, with a relatively smaller gap (57.9% precision and 56.2% recall). In contrast, subject and verb detection have higher recall and similar results as in ParaSCI-arXiv, with an F1-score of 84.2% and 75.6%, respectively.

The stark difference in character and action detection results between the two sources is an interesting pattern, indicating potential over-identification in scientific-themed sentences. This pattern suggests potential areas for refinement and demonstrates the adaptability of the system to different sentence characteristics.

Classification of sentence adherence to CLEARRULES

Shown by Table 4.3, the results for inadherence detection reveal some interesting patterns across different datasets and five inadherence categories. For the ParaSCI-arXiv dataset, AVOIDLONGSUBJ and AVOIDLONGINTRO demonstrate relatively higher F1-scores at 85.7% and 75% respectively, along with a balanced precision and recall. Detecting ALIGNCHARSUBJ inadherence gives a modest 66.7% F1-score. As for the

rest, we get less desirable results, with lower F1-score, particularly for AVOIDINTERRUPTSV at 22.2% and ALIGNACTIONVERB at 48%.

Such a low performance on AVOIDINTERRUPTSV is not reflected very well in the subject and verb span recognition. Thus, this problem may signal a deeper issue rooted in the alignment between the system’s design principles and the author’s conceptual understanding of what constitutes an “interruption.” Future enhancements could focus on refining the semantic definitions within the system.

4.4 General Discussion

The development and evaluation of the CLEARMETRICS system uncover some pitfalls in detecting characters and actions, along with flagging inadherence to clear writing principles. The interconnected nature of these challenges suggests a complex relationship between span detection and inadherence flagging. Understanding this connection is vital and may require a holistic approach to improve the system’s accuracy in assessing adherence to writing principles. Future work should focus on a more nuanced evaluation of alignment rules and possibly integrating more human expertise to resolve these ambiguities.

With that being said, we can safely conclude an inquiry from one of our objectives in this chapter, that there is a potential that abstract meaning representation can be used to accurately detect characters and actions. Different methods for approaching the problem could be used, such as training an end-to-end machine learning model for each task, fine-tuning a pre-trained AMR parsing model, or leveraging a large-language model. Each approach has its own challenge and may require different preparations, such as the need for more labelled data.

Chapter 5

CLEARRULES Evaluation

This chapter presents the detailed methodology of the survey to evaluate CLEARRULES, the result on each writing rule, and a general discussion that ties all observations and insights.

5.1 Methodology

5.1.1 Framework

To answer the first research question, we employ a descriptive-comparative framework because we are not introducing interventions or treatments but rather observing and describing existing conditions. The study aims to describe the clarity of sentences by assessing their adherence to clear-writing principles, specifically the CLEARRULES. For this purpose, participants of the survey are presented with pairs of sentences, each representing adherence to a rule. They were asked to indicate their preference for clearer sentences within each pair. By utilising this comparative approach, we seek to discover participants' natural inclinations towards sentences that exhibit distinct clear-writing principles.

5.1.2 Survey Design

In conducting the survey to evaluate the effectiveness of CLEARRULES, we recognise the potential biases that might influence participants' choices. Therefore, specific measures were taken in the survey design to mitigate these biases:

- **Confirmation bias:** We want to avoid participants' choices being affected by

their preconceived beliefs or expectations about clear writing. To achieve that, we did not inform them about the specific clear writing principles being tested or which sentence in each pair applied them. This approach aimed to ensure that participants evaluated the sentences based on their own judgement of clarity.

- **Comprehension bias:** This bias refers to the tendency for participants to judge the second sentence in a pair as more clear, possibly because they have already processed the information in the first sentence, regardless of its adherence to CLEARRULES. To control this bias, we randomised the appearance of the sentences within each pair. This measure sought to prevent any systematic preference for the second sentence, or even the first if that's the real case, due to the processing effect or simply because of its position in the pair.
- **Pattern bias:** Participants might detect a pattern in the presentation of the questions or the application of CLEARRULES and respond based on that pattern rather than the content of the texts. By randomising the order of the questions, we aimed to disrupt any such patterns and ensure that participants' choices were based on their genuine assessments of the content and their focus at one time is only on one pair of sentences.

While these controls try to ensure the validity of the survey, there might be other unmeasured variables that could introduce additional biases or confounding factors that are not accounted for in the design.

5.1.3 Sentence Selection

The sentences are selected from the train sets of a total of 370 pairs of augmented paraphrase corpora. CLEARFLAGS were predicted using CLEARMETRICS and we choose only pairs whose difference in their CLEARFLAGS is only on one place. For example, if the original sentence has CLEARFLAGS [1, 1, 1, 0, 0] and the paraphrase sentence has [0, 0, 0, 0, 0] (or vice versa), we will not select this pair, because the difference in their CLEARFLAGS occurs at more than one place. Conversely, if the original sentence has CLEARFLAGS [1, 1, 1, 0, 0] and the paraphrase sentence has [1, 1, 0, 0, 0] (or vice versa), we would select this pair for analysis, as the difference is isolated to a single rule adherence, that is the third rule, AVOID-LONGSUBJ. This selection criterion ensures that the comparisons made are focused

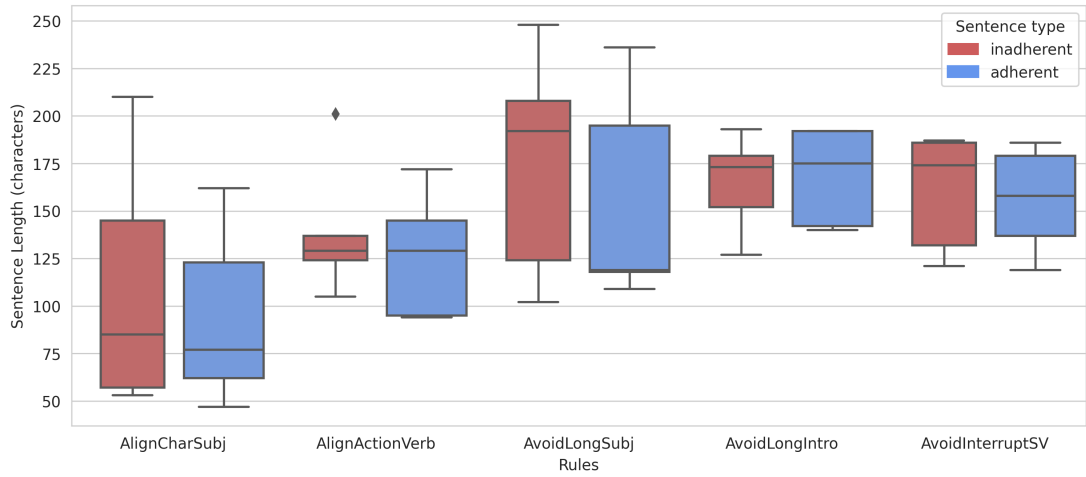


Figure 5.1: Comparison of sentence length distribution between adherent and inadherent sentences for each rule.

on specific differences between the pairs and minimizes the confounding influence of multiple variable changes.

However, from our initial setting, we only get a small pool of candidates, so we run CLEARMETRICS on the rest of MSRP and the first 5000 pairs of ParaSCI-arXiv. Initially, we hoped that the filtering result would be quite straightforward, but some of the pairs cannot be deemed as equal, since sometimes information in one sentence is not presented in another, thus we omitted such pairs. Furthermore, as we can see in Chapter 5, CLEARMETRICS’ prediction on CLEARFLAGS has many false positives. In the end, we had to read the pairs carefully and manually select the correctly predicted and equal pairs.

We prepared five pairs to test for each rule, totalling 25 pairs. The complete list of sentences used in the survey can be read from Appendix D.

We checked character lengths between the adherent and inadherent sets to see if this could be a confounding factor that affected the participants’ preference in the survey later. From Figure 5.1 we can see that the only substantially different average character length is on the AVOIDLONGSUBJ sentence sets. This might be natural because sentences with long subjects tend to use wordy expressions. We do not factor this into the analysis, not only because of the time constraint of the study but also to avoid over-analysing a limited sample size.

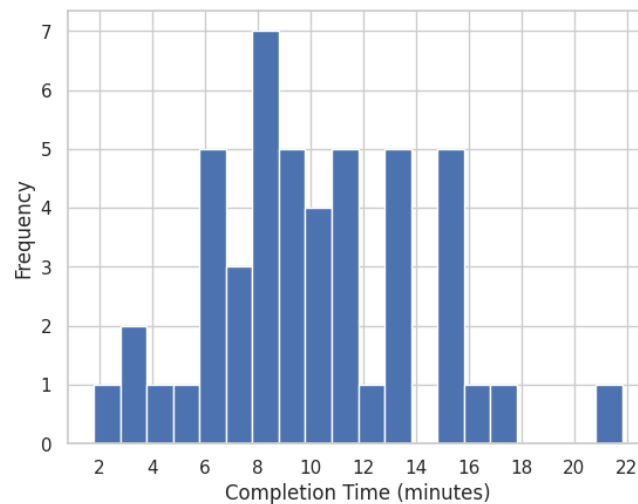


Figure 5.2: Histogram of participants' survey completion time, excluding the right end outliers e.g. responses which taken more than 20 minutes.

5.1.4 Data Collection

The survey was conducted using an online format and was hosted on Qualtrics survey platform¹, which access is provided by the university. This approach allowed for efficient data collection and ensured a uniform presentation of questions to all participants. Despite that, by using an online and asynchronous survey we cannot control the environment of the participants. They might be in distracting settings that could affect their concentration and the quality of their response. We can only ask the participants to condition their environments themselves through an instructive text.

The majority of participants, 44 out of total 50, were recruited from Prolific², a reputable source for academic research participant recruitment. The remaining 6 participants were from our personal connections. All participants have attained at least an undergraduate degree and are fluent in English. Since we are limiting our focus to the context of scientific communication, these criteria were important to ensure participants had a certain level of proficiency and experience with academic or formal writing so that they could effectively evaluate the presented sentences.

We estimated that the survey would take 9 to 15 minutes to complete. However, when we checked the actual survey completion times, we found that some participants fell outside of that range. The longer completion times might have been due to pauses in between questions, while the validity of the shorter ones might be questionable. We

¹<https://edinburgh.eu.qualtrics.com/>

²<https://www.prolific.co/>

decided to exclude responses with completion times below 7 minutes, resulting in a total of 40 valid responses. Still, we will separately analyse both the full sample, A, and the cleaned sample, B.

5.1.5 Data Analysis

We examine the proportions of preferences for each principle to determine if they have a significant impact on enhancing sentence clarity. To assess significance, we conducted multiple hypothesis tests. There are five null hypotheses (H_0) and alternative hypotheses (H_1), which are generalised as follows.

H_0 : Participants are equally likely to choose either sentence, regardless of whether it applies the rule or not.

H_1 : Sentences that apply the rule are more likely to be selected.

In other words, H_0 indicates that there is $p = 0.5$ chance that participants select sentences that apply the rule, e.g. there is no preference between the two sentences. When H_0 is rejected, we can say that H_1 is likely to be true, which implies that the specific CLEARRULES make the sentence clearer. These tests answer our first research question on whether the application of CLEARRULES in scientific-themed sentences improves their clarity.

We adopt the standard initial significance level (α) of 0.05 for each test. Bonferroni correction [Weinstein, 2004] is employed to avoid Type I errors or false positive conclusions, e.g. falsely rejecting the null hypothesis. This makes the significance level drop into $0.05/5 = 0.01$ as the rules are being tested by five questions. We use this corrected significance level, 0.01, in our reporting.

To derive the p -value, we employ the binomial test. The binomial test is used to determine the probability of observing a specific number of successes in a fixed number of independent trials with a constant probability of success. In our context, this test is used to test whether the observed proportion of successes in a sample is significantly different from a hypothesised proportion or a chance level, that is 0.5. This method is appropriate since our survey consists of binary outcomes where participants are selecting one of two sentences. This matches the binomial distribution's representation of 'successes' and 'failures'. To be more specific, we formulate the 'success' event as an instance when a participant selected a CLEARRULES-adherent sentence. Furthermore, our survey design also met the assumption of independent observations and a fixed

number of trials. Each question can be considered an independent trial because the sentence and question order are randomised.

Calculating the probability P of observing exactly k successes in n trials with success probability p is given by the binomial probability formula:

$$P(k; n, p) = \frac{n!}{k!(n-k)!} \times p^k \times (1-p)^{n-k} \quad (5.1)$$

While the calculation for a binomial test is simple, we ease our work by using relevant statistical functions provided by the SciPy package³.

Beyond merely examining statistical significance through p -values, it's important to gauge the magnitude of the difference in participants' preferences. We employed Cohen's h [Cohen, 2013] as a measure of effect size. It will assess the size of the difference in preferences between the adherent and inadherent sets in our survey. By knowing how big that difference is, we can have more insight into the practical significance of the findings. A large value of Cohen's h will indicate a greater difference between the compared groups, which might mean a more substantial preference for one writing style over the other. The h score is defined as the difference between the arcsine transformation of the probabilities of two populations:

$$h = 2(\arcsin \sqrt{p_1} - \arcsin \sqrt{p_2}) \quad (5.2)$$

The measure typically falls within the range of -1 to $+1$. According to the rule-of-thumb [Cohen, 2013], we can interpret the values:

- $h \approx 0$: no effect or no difference between the proportions.
- $h \approx 0.2$: small effect size.
- $h \approx 0.5$: medium effect size.
- $h \approx 0.8$ or higher: large effect size.

Negative values would imply that the direction of the effect is reversed.

5.2 Results and Interpretations

The total of participants we have from all responses is 50, thus we have $n = 5 \times 50 = 250$, referred to as A , to test each writing rule. After excluding the responses whose

³<https://scipy.org/>

completion time is below a threshold, we have 40 responses, meaning $n = 5 \times 40 = 200$, referred to as B, a stricter sample to test each rule. The results are summarised in Table 5.1 for the hypotheses tests and in Figure 5.3 for the effect size measures.

Rule	Probability		p -value		Reject H_0	
	A	B	A	B	A	B
ALIGNCHARSUBJ	0.668	0.662	5.9×10^{-8}	6.5×10^{-8}	T	T
ALIGNACTIONVERB	0.584	0.587	4.7×10^{-3}	5.6×10^{-3}	T	F
AVOIDLONGSUBJ	0.760	0.769	3.2×10^{-17}	1.1×10^{-16}	T	T
AVOIDLONGINTRO	0.528	0.520	0.205	0.297	F	F
AVOIDINTERRUPTSV	0.548	0.560	0.073	0.041	F	F

Table 5.1: Multi-hypotheses tests results with $\alpha = 0.01$. A is the case of the full amount of survey responses ($n = 250$) and B is the case of filtered responses ($n = 200$)

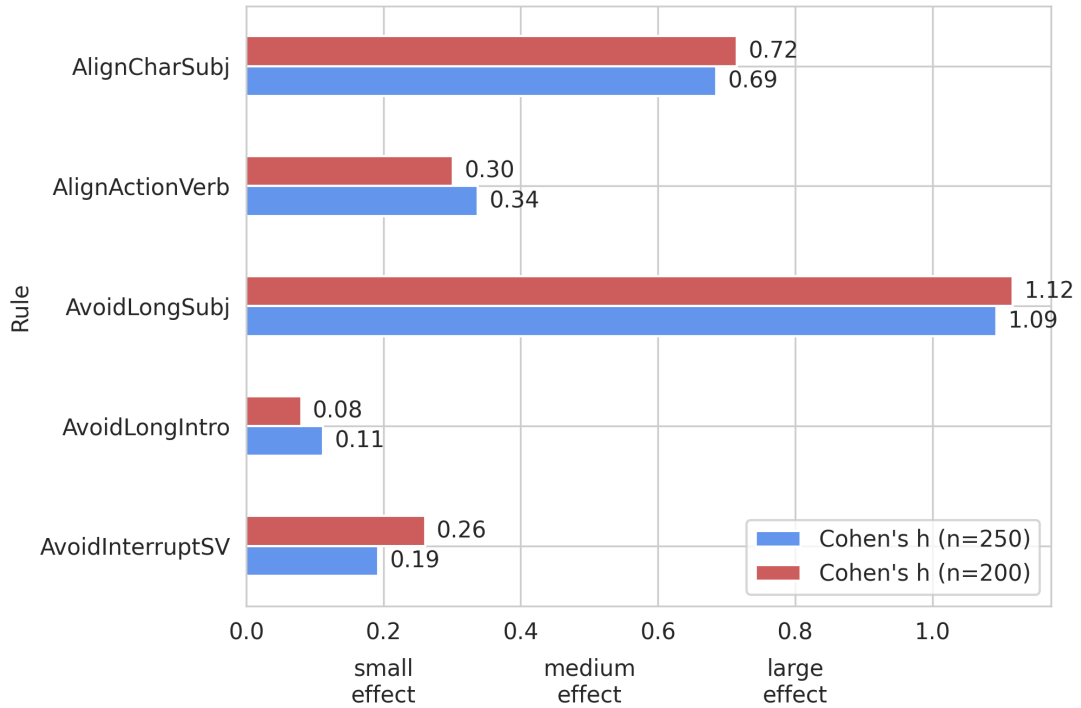


Figure 5.3: Cohen's h for each writing rule. Generally, the effect size of $n = 250$ and $n = 200$ populations across the rules are quite similar.

Firstly, in our evaluation of the adherence to the ALIGNCHARSUBJ rule, both full and cleaned populations exhibited strong preferences for the adherent sentences. The

null hypothesis is rejected and the h measure suggests that ALIGNCHARSUBJ has medium to large effect size for both sample sizes.

The test for ALIGNACTIONVERB revealed a difference in participants' preferences, with a small to medium effect size. While the full population demonstrated statistical significance, this was not observed in the reduced sample. The null hypothesis was rejected for A but not for B, despite B's preference for the adherent sentence being slightly higher.

Further, the results for the AVOIDLONGSUBJ rule are highly significant, both the full population A and the subset B. The effect sizes suggest that the application of this rule in a sentence has a strong effect on the reading experience.

Lastly, the AVOIDLONGINTRO and AVOIDINTERRUPTSV rules did not achieve statistically significant results, failing to reject the null hypotheses. Cohen's h values are relatively low in both populations, signifying a small effect size. This suggests that the long introductory phrases and interruption of the subject-verb sequences only minimally impact the readers.

5.3 General Discussion

The results from the evaluation of CLEARRULES through human surveys offer several insights. Adherence to ALIGNCHARSUBJ and AVOIDLONGSUBJ are significantly appreciated by readers. In contrast, the rule AVOIDLONGINTRO and AVOIDINTERRUPTSV showed minimal impact on readers, indicating that these stylistic elements may be less crucial in scientific writing. Additionally, the mixed results for ALIGNACTIONVERB reveal nuances that could vary with individual preferences or might be subtle enough to be overlooked. This illustrates that some rules may be less pivotal, allowing for more stylistic flexibility.

The potential sensitivity to the sample size is observed in the ALIGNACTIONVERB test. This case manifests one of the general challenges in interpreting p -values close to the significance level. This also raises questions about the generalisability of the findings. Future research may benefit from expanding the sample size.

Chapter 6

Conclusions

This study aims to evaluate the effectiveness of clear-writing principles and automate the process of detecting principles-inadherent sentences within the scientific domain. The following enumerates the contributions of this study, along with the key findings.

1. **Annotation of 740 sentence pairs**, including labels for subjects, verbs, characters, and actions, along with inadherence flags to CLEARRULES.
2. **The CLEARMETRICS system**, achieving F1-score of 74.9% in span detection for ParaSCI-arXiv and 71.7% for MSRP. As for flagging inadherence, the system gives F1-score of 60.2% for ParaSCI-arXiv and 44.% for MSRP.
3. **Survey on CLEARRULES effectiveness**, identifying substantial preferences for rules like aligning subjects with characters and avoiding long subjects, and minimal reader impact for long introductions and interruptions. We discuss the subtlety of alignment between verb-action words and sensitivity to sample size, highlighting complexities in evaluating stylistic elements and the need for more robust methodologies.

We also have highlighted common pitfalls, including over-sensitivity in inadherence flagging and inconsistencies in span detection. Future enhancements could focus on refining algorithms and addressing error propagation. The insights gained also can inform the creation of writing tools, guidelines, and educational programs to elevate the clarity of scientific writing, particularly in paraphrasing practice.

We hope that the lessons learned resonate not only within the corridors of academia but echo in the wider realm of communication, where clarity is not just a virtue but also a necessity.

Bibliography

- [Alvi et al., 2021] Alvi, F., Stevenson, M., and Clough, P. (2021). Paraphrase type identification for plagiarism detection using contexts and word embeddings. *International Journal of Educational Technology in Higher Education*, 18(1):42.
- [Anchiêta and Pardo, 2020] Anchiêta, R. and Pardo, T. (2020). Semantically inspired AMR alignment for the Portuguese language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1595–1600, Online. Association for Computational Linguistics.
- [Artstein, 2017] Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.
- [Banarescu et al., 2013] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking.
- [Bevilacqua et al., 2021] Bevilacqua, M., Blloshmi, R., and Navigli, R. (2021). One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.
- [Chu and Evans, 2021] Chu, J. S. G. and Evans, J. A. (2021). Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118. Citation Key: doi:10.1073/pnas.2021636118
tex.eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.2021636118>.
- [Cohen, 2013] Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Elsevier Science.
- [Dale and Kilgarriff, 2011] Dale, R. and Kilgarriff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Nat-*

- ural Language Generation*, pages 242–249, Nancy, France. Association for Computational Linguistics.
- [Daudaravičius, 2015] Daudaravičius, V. (2015). Automated evaluation of scientific writing: AESW shared task proposal. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–63, Denver, Colorado. Association for Computational Linguistics.
- [Dolan and Brockett, 2005] Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- [Dong et al., 2021] Dong, Q., Wan, X., and Cao, Y. (2021). ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434, Online. Association for Computational Linguistics.
- [Du et al., 2022] Du, W., Raheja, V., Kumar, D., Kim, Z. M., Lopez, M., and Kang, D. (2022). Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.
- [Dwivedi-Yu et al., 2022] Dwivedi-Yu, J., Schick, T., Jiang, Z., Lomeli, M., Lewis, P., Izacard, G., Grave, E., Riedel, S., and Petroni, F. (2022). Editeval: An instruction-based benchmark for text improvements.
- [Gayed et al., 2022] Gayed, J. M., Carlon, M. K. J., Oriola, A. M., and Cross, J. S. (2022). Exploring an ai-based writing assistant’s impact on english language learners. *Computers and Education: Artificial Intelligence*, 3:100055.
- [Gopen and Swan, 1990] Gopen, G. D. and Swan, J. A. (1990). The science of scientific writing. *American scientist*, 78(6):550–558.
- [Greene, 2013] Greene, A. (2013). Writing science in plain english.
- [Grover, 2011] Grover, N. (2011). Scientific writing and communication: Papers, proposals, and presentations. *Biochemistry and Molecular Biology Education*, 39(2):181–181.

- [Gunning, 1969] Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13. Citation Key: doi:10.1177/002194366900600202 tex.eprint: <https://doi.org/10.1177/002194366900600202>.
- [Heard, 2014] Heard, S. B. (2014). On whimsy, jokes, and beauty: can scientific writing be enjoyed? *Ideas in Ecology and Evolution*, 7(1).
- [Honnibal and Johnson, 2015] Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- [Huang et al., 2023] Huang, K.-H., Iyer, V., Hsu, I.-H., Kumar, A., Chang, K.-W., and Galstyan, A. (2023). ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.
- [Issa et al., 2018] Issa, F., Damonte, M., Cohen, S. B., Yan, X., and Chang, Y. (2018). Abstract Meaning Representation for paraphrase detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 442–452, New Orleans, Louisiana. Association for Computational Linguistics.
- [Ito et al., 2020] Ito, T., Kuribayashi, T., Hidaka, M., Suzuki, J., and Inui, K. (2020). Langsmith: An interactive academic text revision system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 216–226, Online. Association for Computational Linguistics.
- [Jurafsky and Martin, 2023] Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*, chapter 18. 3 (draft) edition.
- [Kim et al., 2022] Kim, Z. M., Du, W., Raheja, V., Kumar, D., and Kang, D. (2022). Improving iterative text revision by learning where to edit from other revision tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9986–9999, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- [Kincaid et al., 1975] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida.
- [Kovatchev et al., 2018] Kovatchev, V., Martí, M. A., and Salamó, M. (2018). ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- [Liao et al., 2018] Liao, K., Lebanoff, L., and Liu, F. (2018). Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- [Liu et al., 2018] Liu, Y., Che, W., Zheng, B., Qin, B., and Liu, T. (2018). An AMR aligner tuned by transition-based parser. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2430, Brussels, Belgium. Association for Computational Linguistics.
- [Lo et al., 2020] Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (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.
- [Martínez Lorenzo et al., 2023] Martínez Lorenzo, A. C., Huguet Cabot, P. L., and Navigli, R. (2023). Cross-lingual AMR aligner: Paying attention to cross-attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1726–1742, Toronto, Canada. Association for Computational Linguistics.
- [Matthews and Matthews, 2014] Matthews, J. R. and Matthews, R. W. (2014). *Successful Scientific Writing: A Step-by-Step Guide for the Biological and Medical Sciences*. Cambridge University Press, 4 edition.
- [Miller, 1956] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

- [Montgomery, 2017] Montgomery, S. L. (2017). *The Chicago guide to communicating science*. University of Chicago Press.
- [Ng et al., 2013] Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.
- [Nivre et al., 2017] Nivre, J., Zeman, D., Ginter, F., and Tyers, F. (2017). Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.
- [Oral and Eryiğit, 2022] Oral, K. E. and Eryiğit, G. (2022). AMR alignment for morphologically-rich and pro-drop languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 143–152, Dublin, Ireland. Association for Computational Linguistics.
- [Palmer et al., 2005] Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106. Citation Key: 10.1162/0891201053630264 tex.eprint: <https://direct.mit.edu/coli/article-pdf/31/1/71/1798172/0891201053630264.pdf>.
- [Senter and Smith, 1967] Senter, R. and Smith, E. A. (1967). Automated readability index. Technical report, Technical report, DTIC document.
- [Vila et al., 2014] Vila, M., Martí, M. A., Rodríguez, H., et al. (2014). Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.
- [Weisstein, 2004] Weisstein, E. W. (2004). Bonferroni correction. <https://mathworld.wolfram.com/>.
- [Wieting and Gimpel, 2018] Wieting, J. and Gimpel, K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

[Williams and Bizup, 2013] Williams, J. and Bizup, J. (2013). *Style: Lessons in clarity and grace*. Pearson Education.

Appendix A

Ten Principles for Writing Clearly

1. Distinguish real grammatical rules from folklore.
2. Use subjects to name the characters in your story.
3. Use verbs to name their important actions.
4. Open your sentences with familiar units of information.
5. Get to the main verb quickly:
 - Avoid long introductory phrases and clauses.
 - Avoid long abstract subjects.
 - Avoid interrupting the subject-verb connection.
6. Push new, complex units of information to the end of the sentence.
7. Begin sentences that form a unit with consistent subjects/topics.
8. Be concise:
 - Cut meaningless and repeated words and obvious implications.
 - Put the meaning of phrase into one or two words.
 - Prefer affirmative sentences to negative ones.
9. Control sprawl:
 - Don't tack more than one subordinate clause onto another.
 - Extend a sentence with resumptive, summative, and free modifiers.
 - Extend a sentence with coordinate structures after verbs.
10. Above all, write to others as you would have others write to you.

Appendix B

Annotation Guidelines

Span Labelling

1. One sentence is presented to work on at one time.
2. Label the word span of the corresponding subject, including:
 - Noun or pronoun subjects, e.g., *Mary*, *cat*, and *they*.
 - Gerund subjects, e.g., *running* in *Running is her hobby*.
 - Infinitive subjects, e.g., *to write* in *To write is to practice thinking*.
 - Clause subjects, e.g., *that the experiment failed* in *That the experiment failed is no surprise*.
 - Compound subjects, e.g., *the wolf and the sheep*.
 - Expletive subjects or dummy subjects, e.g., *there* in *There are two approaches*.
 - Nominalised subjects, e.g., *the interference*.
 - Phrase subject, e.g., *the sample in the left group* in *The sample in the left group is bigger*.
3. Label the word span of the main verb phrase. This includes:
 - Lexical verbs, e.g., *help*, *know*, and *appear*.
 - Auxiliary verbs, e.g., *be*, *have*, and *will*.
 - Phrasal verbs, e.g., *break into* and *wipe out*.

4. There could be multiple subjects and verbs, limit each into two instance. Such sentence usually contains quoted speech, e.g., *"We will review the plan," the president said.*
5. Label the actions, which could in form of verb, adjective, or noun.
6. Label the characters corresponding characters, which could be flesh-and-blood characters or some abstract ones.
7. Link the subjects to the corresponding verbs and the characters to the corresponding actions.

CLEARRULES Inadherence Classification

1. Check whether one of the character spans overlaps with one of the subject spans. The there is no overlap, then label the text as "Character-Subject Misalignment."
2. Check whether one of the action spans overlaps with one of the verb spans. The there is no overlap, then label the text as "Action-Verb Misalignment."
3. Count the words for each subject span. If each of them contains more than 8 counts, mark the text as having "Long Abstract Subject."
4. Check if the sentence has introductory phrases and clauses. If it has, count the words in that intro. If it exceeds more than 8 counts, mark the text as having "Long Introductory Phrases and Clauses."
5. Check if the main subject and verb are interrupted by some words, for example, a clause. If that interruption exceeds 4 words, tick the "Subject-Verb Interruption" option.

Note that in counting words, the span of an entity is regarded as only a count and punctuation is ignored.

Appendix C

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.

Ethics application number: 299455

Date when approval was obtained: 2023-07-20

The participants' information sheet and a consent form are included in Appendix D.

Appendix D

Survey Questionnaire

Figures after the table are the screenshots of the survey, page by page.

Table D.1: List of sentences presented to the participants.

#	Rule	Original sentence	Paraphrase sentence
1	1	The fact that she admitted guilt impressed the committee.	Her admission of guilt impressed the committee.
2	1	Contradictions among the data require an explanation.	We were required to explain the contradictions among the data.
3	1	The ResNet50 pretrained on the ImageNet dataset is employed as our initialized model.	We make use of the ResNet-50 architecture pretrained on the ImageNet dataset.
4	1	Edge computing is an emerging computing paradigm that is transforming the landscape of provision and consumption of computing services for a wide range of applications and end users at the edge of the internet.	Edge computing is an emerging paradigm which proposes to move cloud services closer to the users and to the devices that produce data, at the edge of the network.

Table D.1: List of list of sentences presented to the participants. (cont.)

#	Rule	Original sentence	Paraphrase sentence
5	1	The exchange-correlation energy was evaluated with the help of the Perdew-Burke-Erzenhof approach, within the generalised gradient approximation.	The generalized gradient approximation of Perdew, Burke and Ernzerhof was used to describe the exchange-correlation energy.
6	2	Our more effective presentation of our study resulted in our success, despite an earlier start by others.	Although others started earlier, we succeeded because we presented our study more effectively.
7	2	A nominalization is a replacement of a verb by a noun, often resulting in displacement of characters from subjects by nouns.	When a nominalization replaces a verb with a noun, it often displaces characters from subjects.
8	2	The adaptive construction method works well for reducing correlations between the sampled data as shown by the numerical results.	The numerical results show that the adaptive construction method significantly reduces the correlations between the sampled data.
9	2	A large number of domain-specific network programming languages have been proposed over the past few years, driven by rapidly expanding infrastructures and the emergence of software-defined networking.	The recent emergence of software-defined networking has led to the development of a number of domain-specific programming languages for networks.

Table D.1: List of list of sentences presented to the participants. (cont.)

#	Rule	Original sentence	Paraphrase sentence
10	2	The aim of this paper has been to solve the geodesic equation exactly and present the complete set of solutions to the geodesic equation.	In this paper, we are aiming at solving the geodesic equation exactly by using numerical techniques and at exploring the complete set of solutions of the geodesic equation.
11	3	Research demonstrating the soundness of our reasoning and the need for action supported this decision.	This decision was supported by research demonstrating the soundness of our reasoning and the need for action.
12	3	A decision about forcibly administering medication in an emergency room setting despite the inability of an irrational patient to provide legal consent is usually an on-scene medical decision.	Medical professionals usually decide on-scene whether to forcibly medicate patients who are unable to legally consent.
13	3	Bayesian networks or graphical models based on directed acyclic graphs are widely used to model complex causal systems arising from a variety of research areas, including computational biology, epidemiology, sociology, and environmental management.	Directed acyclic graph models, also known as Bayesian networks, are widely used to model causal relationships in complex systems across various fields such as computational biology, epidemiology, sociology, and environmental management.
14	3	Much of the work with deep learning in natural language processing has involved the learning of word vector representations.	In natural language processing, deep learning methods have been mostly focused on learning word vector representations.

Table D.1: List of list of sentences presented to the participants. (cont.)

#	Rule	Original sentence	Paraphrase sentence
15	3	An algorithm very commonly used in practice to determine the total space of possible resultant wrenches as long as each individual contact force obeys friction constraints was introduced by Ferrari and Canny.	Ferrari and Canny introduced a very efficient geometric method for determining the total space of possible resultant wrenches as long as each individual contact wrench obeys friction constraints.
16	4	When a company focuses on hiring the best personnel and then trains them not just for the work they are hired to do but for higher-level jobs, it is likely to earn the loyalty of its employees.	A company is likely to earn the loyalty of its employees when it focuses on hiring the best personnel and then trains them not just for the work they are hired to do but for higher-level jobs.
17	4	When patients appear in a Trauma Center and behave so irrationally that they cannot legally consent to treatment, only the attending physician can decide whether to medicate them.	It is only the attending physician who can decide whether to medicate patients that cannot legally consent to treatment in the Trauma Center due to their irrational behaviour.
18	4	With the help of powerful well-designed deep neural networks, great progresses have been made in the field of object detection.	In recent years, the accuracy of object detection has been dramatically improved thanks to the advance of deep convolutional neural network.

Table D.1: List of list of sentences presented to the participants. (cont.)

#	Rule	Original sentence	Paraphrase sentence
19	4	While models following this paradigm have been found very useful in a number of natural language processing tasks, they do not scale up to the level of phrases or sentences.	Models following this paradigm have been proved useful in many natural language processing tasks, but in general they do not scale up to larger text constituents such as phrases and sentences.
20	4	In the special case when v arises from a time-invariant control system, the complexity of our algorithm agrees with that of the well-known GS algorithm.	Our algorithm therefore performs as well as the GS algorithm in the special case when v happens to arise from a time-invariant control system.
21	5	We must develop, if we are to become competitive with other companies in our region, a core of knowledge regarding the state of the art in effective industrial organizations.	If we are to compete with other companies in our region, we must develop a core of knowledge about the state of the art in effective industrial organizations.
22	5	Some scientists, because they write in a style that is impersonal and abstract, do not easily communicate with laypeople.	Some scientists do not easily communicate with laypeople because they write in a style that is impersonal and abstract.
23	5	The relatively recent innovation of synthetic materials, made possible with ultra-cold atomic gases, has added a vitally important tool to study many-body physics in experimental systems.	Recent experimental progress with cold atomic gases has opened the door for realizing quantum many-body systems and simulating some of the most useful models of many-body physics.

Table D.1: List of list of sentences presented to the participants. (cont.)

#	Rule	Original sentence	Paraphrase sentence
24	5	Metasurfaces, which are the two-dimensional counterparts of volume metamaterials, have attracted much attention over the past years.	Presented as two-dimensional equivalents of volumetric metamaterials, metasurfaces have attracted a significant interest in recent years.
25	5	Cognitive radio, with its capability to flexibly configure its transmission parameters, has emerged in recent years as a promising paradigm to enable more efficient spectrum utilization.	With its flexibility in configuring the transmission parameters, cognitive radio has attracted intensive research in recent years because of pressing demand for efficient spectrum usage.

Investigating the Effectiveness of Clear-Writing Principles for Scientific Communication

I am [REDACTED], a student at the University of Edinburgh, MSc Speech and Language Processing. Thank you for having an interest to participate in my study. Before giving consent, please take some time to read the information to understand your rights and how you will participate.

This study was certified according to the Informatics Research Ethics Process, reference number 299455.

What is the purpose of the study?

The study aims to test the effectiveness of writing principles proposed by Williams (1981) on improving the clarity and readability of sentences written for scientific publications. This study will help future research in natural language processing, especially for paraphrasing task, to be more attuned to the complexity of writing scientific contents.

Why have I been asked to take part?

The research target group is people who have at least obtained a bachelor's degree. A bachelor's degree attainment is a proxy to ensure that the participants are familiar with the English language and scientific-themed content.

Do I have to take part?

No – participation in this study is entirely up to you. But because we don't record any personal information, it is hard to withdraw from the study once a response is submitted. If you want to participate but still have concerns, please reach me through [REDACTED]

What will happen if I decide to take part?

You will be given a task to compare 25 pairs of sentences. You have to decide which sentence gives clearer meaning. The domains of the content are mainly computer science and physics, with some business and medicine.

The questionnaire is designed to be completed in 10 to 13 minutes. You can pause the survey and continue later as long as the same browser is used.

Will I be compensated?

You are eligible to be paid up to £2 (or IDR 30k) after completing your participation in this study. A compensation request is separated from the submission of the questionnaire to protect anonymity of the responses. You will have to send me an email of **the completion code** which you will find after you completed the form.

If you are coming from **Prolific**, you will be compensated accordingly through the platform. Please submit **the completion code** after you send the form.

If you prefer to received no compensation, you don't have to do anything after submitting a response.

Figure D.1: The introduction to the questionnaire (1/2), a substitute for Participant Information Sheet. The researcher's information is omitted for this dissertation's anonymity.

Are there any risks associated with taking part?

No. The content of this survey has been curated to avoid any risk to the participants.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Your response can also be used for future research. Your data may be archived for a maximum of four years.

What about the data protection and confidentiality?

Your data will be processed in accordance with the UK's [Data Protection Law](#). Your response will be referred to by a unique participant number. The data will only be viewed by me, Ilma, the researcher. All electronic data will be stored on a password-protected encrypted computer.

For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact me at [REDACTED]

If you wish to complain about the study, please contact inf-ethics@inf.ed.ac.uk, providing the study title and details of the nature of your complaint.

Consent

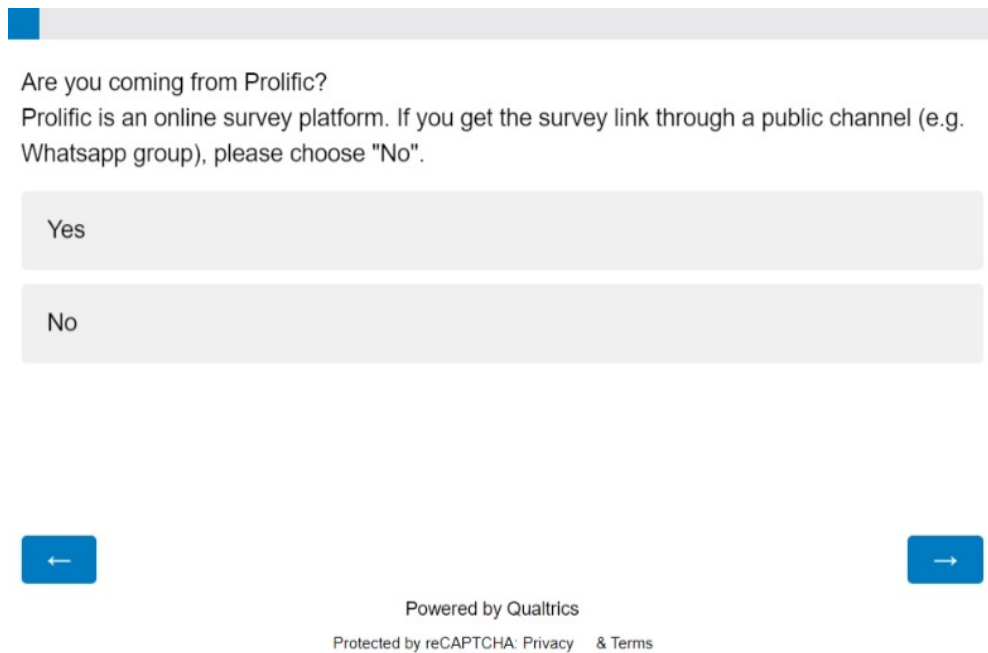
To proceed, you have to agree with the following statements:

1. I have read and understood the above information.
2. I understand that my participation is voluntary, and I can withdraw at any time.
3. I consent to my anonymised data being used in academic publications and presentations.
4. I allow my data to be used in future ethically approved research.
5. I will follow the survey instructions with the best of my ability

By activating this button, I agree with the above statements.



Figure D.2: The introduction to the questionnaire (2/2), a substitute for Participant Information Sheet. The researcher's information is omitted for this dissertation's anonymity.



Are you coming from Prolific?
Prolific is an online survey platform. If you get the survey link through a public channel (e.g. Whatsapp group), please choose "No".

Yes

No

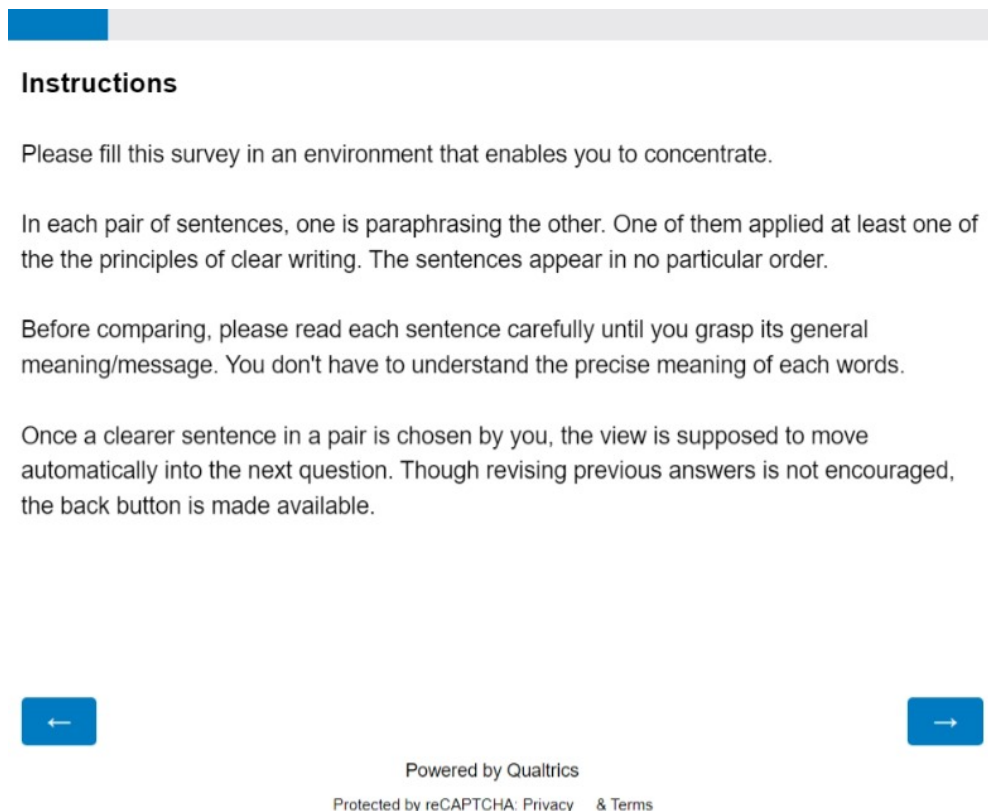
←

→

Powered by Qualtrics

Protected by reCAPTCHA: Privacy & Terms

Figure D.3: A screen of question that asks the source of the participant.



Instructions

Please fill this survey in an environment that enables you to concentrate.

In each pair of sentences, one is paraphrasing the other. One of them applied at least one of the the principles of clear writing. The sentences appear in no particular order.

Before comparing, please read each sentence carefully until you grasp its general meaning/message. You don't have to understand the precise meaning of each words.

Once a clearer sentence in a pair is chosen by you, the view is supposed to move automatically into the next question. Though revising previous answers is not encouraged, the back button is made available.

←

→

Powered by Qualtrics

Protected by reCAPTCHA: Privacy & Terms

Figure D.4: The instructions for participants.

Choose the sentence that delivers more clarity and easier to digest.

Much of the work with deep learning in natural language processing has involved the learning of word vector representations.

In natural language processing, deep learning methods have been mostly focused on learning word vector representations.



Powered by Qualtrics

Protected by reCAPTCHA: Privacy & Terms

Figure D.5: An example of the question.

Now that you have submitted your response, you are eligible for pay. Below is the code needed to obtain compensation for your complete response. Please copy it or screenshot the screen. **A compensation request without the code will not be recognised.**

CAGVLDWW

If you are coming from Prolific, use the code to acquire the pay. Otherwise, contact me at i.a.fiddien@sms.ed.ac.uk with [Survey completed] as the header. Please include the code above and your bank details in the body of the email. Note that I can only pay through a UK or Indonesia-based bank.

Thank you.

Powered by Qualtrics

Protected by reCAPTCHA: Privacy & Terms

Figure D.6: Closing page of the questionnaire.

Appendix E

Data for Visualisations

	Character	Action	Subject	Verb
StyleExamples	226	347	140	140
ParaSCI-arXiv-train	238	360	301	316
ParaSCI-arXiv-test	81	86	93	91
MSRP-train	315	413	243	242
MSRP-test	159	169	128	127

Table E.1: The data used to produce Figure 3.3.

	ALIGNCHARSUBJ	ALIGNACTIONVERB	AVOIDLONGSUBJ	AVOIDLONGINTRO	AVOIDINTERRUPTSV
StyleExamples	45	10	4	3	3
ParaSCI-arXiv-train	10	17	13	10	3
ParaSCI-arXiv-test	13	0	6	4	2
MSRP-train	11	1	6	5	13
MSRP-test	9	9	1	1	5

Table E.2: The data used to produce Figure 3.4.

Table E.3: The complete data used to produce Figure 5.2: Survey completion time.

Participant Number	Duration (seconds)
1	220
2	108
3	492
4	251
5	452
6	525
7	301
8	502
9	388
10	539
11	389
12	581
13	443
14	575
15	451
16	603
17	636
18	781
19	621
20	671
21	661
22	913
23	993
24	910
25	924
26	810
27	944
28	902
29	1062
30	1310
31	594
32	688

Table E.3: The complete data used to produce Figure 5.2: Survey completion time.
(cont.)

Participant Number	Duration (seconds)
33	745
34	2098
35	10534
36	818
37	169
38	361
39	350
40	471
41	574
42	484
43	513
44	513
45	380
46	551
47	696
48	704
49	776
50	818

	Cohen's h (n=250)	Cohen's h (n=200)
Rule		
AlignCharSubj	0.685334	0.715142
AlignActionVerb	0.337601	0.301137
AvoidLongSubj	1.093702	1.117201
AvoidLongIntro	0.112059	0.080021
AvoidInterruptSV	0.192296	0.260738

Table E.4: The data used to produce Figure 5.3: Effect size of each writing rule