Best practices for scientific writing and editing

Luc Rocher

19th October 2024

This document lists structured rules for effective *reader-centric writing* in computational social science and human-computer interaction articles. It serves as a practical checklist for editing and reviewing manuscripts.

1 General rules

- Break down complex ideas into simpler, more digestible parts.
- Lead the reader from what they know to what they don't know.
- Your writing style should not get in the way of your work.
- Every word serves a purpose.

Think conservatively about the reader.

- The reader does not read linearly but zaps to find the most interesting parts of your work.
- The reader has a very short attention span and memory loss.
- The reader is initially on your side; avoid antagonising them.

2 Refining the narrative

- Aim for a single, simple narrative.
- Do not follow a chronological 'biography' of your research progress that led to this manuscript.
- The highlighted issues that led to your solutions should be equally commensurable.
- Always clearly separate past literature and your own work, without back-and-forth.
- An article is not a murder mystery or a puzzle. Present the structure of the argument before the reader can digest it.
  - ✗ *"Here, we explore various aspects of algorithmic bias and its implications. We conducted several experiments and analyses, which led to some interesting findings which we present in Section 7."*
  - ✓ *"Here, we examine algorithmic bias in facial recognition systems. We analyze the performance disparities across different demographic groups. We then investigate the root causes of these disparities, and show them to arise from* [...]*."*

3 Paragraph structure and signposting

- One paragraph = one idea.

  For example, instead of explaining a ML algorithm in one dense paragraph, break it down into steps: data collection, feature extraction, model training, and prediction. To help you, try summarising each paragraph in one sentence; if you can't, it might contain multiple ideas.

- Begin each paragraph with a short sentence that introduces or summarises its content.

  A section should be roughly understood by reading only the first sentence of each paragraph.

- Use headings and subheadings to guide the reader.

  Limit the use of subheadings for journals; use them extensively for conferences.

- In long articles, begin a section with a short paragraph or sentence signposting its overall content.

  This is particularly true for longer conference articles. Example: "In previous sections, we have analyzed the impact of external factors such as the SDG algorithm, synthetic data size, and differential privacy defenses on the success of our attack. Here, we further analyze the impact of the internal parameters used to define the attack and provide an explanation for why our attack works. [...]"

- Start sentences and paragraphs from what the reader knows before introducing new information, facts, or terminology.

  Begin with easier information or information the reader is familiar with, and end with newer or more complex information.

- Aim for five to seven sentences per paragraph maximum.

4 Transitions and connectives

- Ensure effective transitions between sections, between paragraphs, and between sentences.

  A successful transition between two paragraphs should take place in the first sentence of the second paragraph. Do not use the last sentence of a paragraph to start a transition, the reader might not read it.

- Use transitional words sparingly.

  To guide readers, use transitional words (e.g., in contrast, yet, however, in addition, similarly, nevertheless, etc.) at the beginning of paragraphs that mark a sharp transition only. Do not use them for all paragraphs.

- Avoid overuse of the same transitional words.

5 Parallel structure

- Use parallel structure in lists and series, across sentences but also paragraphs.
  - ✗ *"Our study aims to identify key user behaviours, analysing interaction patterns, and we will propose design improvements."*
  - ✓ *"Our study aims to identify key user behaviours, analyse interaction patterns, and propose design improvements."*
- Use parallelism in headings and subheadings.
  - ✗ *"4.1 Collecting User Data, 4.2 Analysis of User Behaviour, 4.3 Design Changes Implementation"*
  - ✓ *"4.1 Collecting User Data, 4.2 Analyzing User Behavior, 4.3 Implementing Design Changes"*
- Use parallel structure in comparative statements.
  - ✗ *"Older participants were less likely to share personal information online, while younger users frequently disclosed private details."*
  - ✓ *"Older participants were less likely to share personal information online, while younger participants were more likely to share private details."*

6 Language and style

6.1 Verb tense consistency

- Use present tense for general truth and analyses.
- Use the present perfect tense when referring to previous or ongoing research.
- Use simple past tense for completed actions, e.g., in Data Collection and Discussion.

6.2 Hedging and qualifying statements

- Use appropriate and consistent hedging language for uncertainties.
- Avoid vagueness or overgeneralisation and use specific, relevant qualifiers to limit the scope of your statements.
  - ✗ *"All users seem to prefer dark mode interfaces."*
  - ✓ *"In our study, the majority of participants expressed a preference for dark mode interfaces."*
- Balance assertiveness with academic caution.
  - ✗ *"Our results prove that social media causes depression."*
  - ✓ *"Our findings suggest a link between social media use and depressive symptoms but do not establish causality."*
- Be specific and qualify broad statements.
  - ✗ *"Machine learning algorithms are more efficient than traditional methods."*

✓ *"In tasks involving large datasets with complex patterns, machine learning algorithms often demonstrate higher efficiency compared to traditional methods."*

6.3 Academic tone

- Use confident language and avoid hedging statements unnecessarily.

  ✗ *"This study may provide insights into user behaviour,"*
  ✓ *"This study provides insights into user behaviour."*

- Avoid colloquialism and slang.

- Avoid emotional language or overly dramatic statements.

  ✗ *"Our groundbreaking study revolutionizes the field of FAccT."*
  ✓ *"Our study contributes new insights to FAccT research by demonstrating…"*

- Use inclusive language, ensure clear antecedents for all pronouns, and avoid ambiguous pronoun references.

  Use "they" instead of "he/she" or "she" (often seen in CS papers) when referring to an unknown person.

6.4 Terminology and definitions

- Avoid jargon and explain technical terms when necessary.

- Spell out acronyms at least once per section.

- Explain statistical terms in simple English terms.

  This also applies when defining a term with an equation. If the statistical term is not common in the field, cite the original authors or cite a relevant textbook.

- Maintain consistent terminology throughout the paper.

  Using different terms for the same idea will confuse the reader. For example, do not mix "our method", "our framework", and "our approach" interchangeably. This also applies to verbs such as, e.g., "show", "examine", or "investigate".

- 'Data' is plural.

  Instead of "data was collected", write "data were collected".

6.5 Clarity and concision

- Use active voice for readability.

  ✗ *"It was found that the interface was preferred by users."*
  ✓ *"Users preferred the new interface."*

- Eliminate redundant or unnecessary words and phrases.

  ✗ *"It is interesting to note that the survey revealed…"*
  ✓ *"The survey revealed…"*

- Avoid overly long or convoluted sentences.
- Use verbs to describe action, state, or occurrence. Do not use nouns.
  - ✗ *"We performed an analysis of the impact of data representation on model fairness."*
  - ✓ *"We analysed how data representation impacts model fairness."*
- Avoid awkward-sounding possessives by restructuring phrases or using attributive nouns.
  - ✗ *"the social network's users' behaviours"*
  - ✓ *"user behaviours in the social network"*
- Avoid split infinitives by placing adverbs after the full infinitive.
  - ✗ *"to deeply analyse the data"*
  - ✓ *"to analyse the data in depth"*
- Use concrete examples to illustrate abstract concepts.
- Avoid metaphors and analogies that are not explicit nor relevant.
  - ✗ *"Our machine learning model acts like a skilled detective, sifting through the haystack of data to find the needle of insight."*
  - ✓ *"Our machine learning model analyzes large datasets to identify specific temporal patterns."*

7 Formatting and conventions

7.1 Numbers and units

- Spell out single-digit numbers used as adjectives. (e.g., "five participants").
- Use numerals with units of measure (e.g., "5 cm").
- Spell out numbers at the start of a sentence (e.g., "Fifteen users completed the survey").
- Use numerals for double-digit numbers not at the start of a sentence (e.g., "The study included 23 participants").

7.2 Punctuation and typography

- Use hyphens to clarify the relationship between words.

  "Low temperature impact" (without a hyphen) suggests a low impact of the temperature, whereas "low-temperature impact" (with a hyphen) suggests the impact of or at low temperature. [1]

- Use the serial Oxford comma.

  "We collected data from surveys, interviews, and social media platforms."

- Use the hyphen (`-`) for compound words, the en-dash (`--`) for ranges, and the em-dash (`---`) for parenthetic statements.

1: Example from Nature English Communication for Scientists.

"The user-centered approach—which emphasises iterative prototyping—led to a 30–40% improvement in completion rates."

7.3 Visual elements: figures, tables, diagrams, photographs, maps

- Write clear and informative captions.
- Ensure captions can stand alone.
- Reference all panels and insets in a figure's caption. Systematically use labels (a), (b), (c), etc.
- Spell out all abbreviations and explain symbols used.
- Ensure that all figures and tables are referenced at least once in the main text.
- Maintain consistency in length, style, and format of captions within each type of visual element.

  It is acceptable to have different caption styles for figures vs. tables, but be consistent within each type. You might e.g., use brief, one-sentence captions for tables and longer, paragraph-style captions for figures.
- Use a consistent format for referencing figures and tables in the text.

  For figures, choose either "Fig." or "Figure" for in-text references and do not mix. For tables, choose either "Table" or "Tbl." consistently. Some publishers have rules with different formats in the main text vs. captions (e.g., "Fig." in text, "Figure" in captions).
- Use a prefix for visual elements in appendix ("Fig. A1") and Supplementary Information ("Fig. S1") if the publisher allows.
- Avoid rasterized images when vector images can be used. If rasterized, ensure at least 300dpi at print size.

If unsure, check on `https://pixelcalculator.com`

Caption for a figure with two side-by-side panels and one inset

The model predicts correct re-identifications with high confidence. (a) Receiver operating characteristic (ROC) curves for USA populations (light ROC curve for each population and a solid line for the average ROC curve). Our method accurately predicts the (binary) individual uniqueness. (Inset) False-discovery rate (FDR) for individual records classified with $\xi > 0.9$, $\xi > 0.95$, and $\xi > 0.99$. For re-identifications that the model predicts are likely to be correct, only 5.26% of them are incorrect (FDR). (b) Our model outperforms by 39% the best theoretically achievable prediction using population uniqueness across every corpus. A red point shows the Brier Score obtained by our model, when trained on a 1% sample. The solid line represents the lowest Brier Score achievable when using the exact population uniqueness while the dashed line represents the Brier Score of a random guess prediction (BS = 1/3).

7.4 Citations and references

- Scientific writing is evidenced-based. It is better to have too many references than too few.
- Use author names as the subject of sentences rather than citations.

  ✗ *"A recent study by [11] showed…"*
  ✓ *"A recent study by Patel et al. showed… [11]."*

- Place reference numbers at the end of a sentence or clause, preceding the punctuation mark.

  "Machine learning algorithms have revolutionised natural language processing [6], leading to breakthroughs in sentiment analysis [7,8]."

- Ensure quotation marks are consistent (single vs. double), matched between opening and closing marks, and consistent with punctuation.

  Some publishers have specific rules, by default use quotations to reproduce words verbatim from prior work, not for general concepts.

- Proofread your references and fill out all missing fields until there is no LaTeX warning. Ensure that you are not using out-of-date preprints instead of published articles.

8 Specific article sections

8.1 Abstract

- The abstract should be easily understood by laypeople and the general public.
- Follow the Nature abstract structure.

  Basic introduction. More detailed background. General problem. Here we show. Main result. Reframing into general context. Broader perspective.

- Limit factual findings such as numbers to one or two.
- Always avoid acronyms in the abstract.

Example of a short abstract for Nature Comms. (200w max)

While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we

find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

8.2 Introduction

- The introduction is the most important part of an article, and should be easily understood by a journalist.
- The introduction should form a self-contained mini paper.

  By default, include: (1) broad context, (2) specific context, (3) motivation/cue, (4) research problem, (5) contribution, and (6) broader impact. You may want to remove (1) or (6).
- If short on space, it is okay to open with a single paragraph that provides context, motivation, and summarises the contribution.

  Using sentences such as "This article uses XX method, applied to YY data, to answer ZZ question."
- Focus on addressing significant, established problems rather than 'research gaps'.

  A lack of research on a topic doesn't automatically make it a worthy research question. A time-tested introduction instead consists of identifying crucial, well-documented problems in the field that remain unsolved.
- Present your work as building upon, rather than criticising, past research.

  Avoid aggressive statements towards past research, especially in the Introduction. Often, you can avoid mentioning past research in Introduction and expand on the its limitations in the Discussion section only.

Example of a complete Introduction structure

In the last decade, the ability to collect and store personal data has exploded. [...]

However, the large-scale collection and use of detailed individual-level data raise legitimate privacy concerns. [...]

De-identification, the process of anonymizing datasets before sharing them, has been the main paradigm used in research and elsewhere to share data while preserving people's privacy. [...]

Yet numerous supposedly anonymous datasets have recently been released and re-identified. [...]

Statistical disclosure control researchers and some companies are disputing the validity of these re-identifications: as datasets are always incomplete, journalists and researchers can never be sure they have re-identified the right person even if they found a match. [...]

[One paragraph with a practical example]

Our paper shows how the likelihood of a specific individual to have been correctly re-identified can be estimated with high accuracy even when the anonymized dataset is heavily incomplete. [...]

8.3 Results

- Structure Results in small self-contained units of work.

  Present results in independent focused units, typically as single paragraphs or small groups of 2–3 paragraphs with subheadings. The reader expects to verify that your conclusions are backed by evidence. However, in mixed-methods studies, a more integrated structure might be appropriate.

- Dedicate one paragraph per figure or table, without overlap.

  Structure these paragraphs as follows: (a) begin with 'Figure X shows…' or 'Table Y shows…', (b) Provide 2-3 sentences describing key data or descriptive statistics from broad ones to specific ones, (c) conclude with 1-2 sentences highlighting main takeaways.

- If a figure includes multiple panels, dedicate one paragraph per panel—unless the panels are closely related.

Example of a Results paragraph

Figure 2a shows that, when trained on 1% of the USA populations, our model predicts very well individual uniqueness, achieving a mean AUC (area under the receiver-operator characteristic curve (ROC)) of 0.89.[1] For each population, to avoid overfitting, we train the model on a single 1% sample, then select 1000 records, independent from the training sample, to test the model.[2] For re-identifications that the model predicts to be always correct (estimated individual uniqueness >95%), the likelihood of them to be incorrect (false-discovery rate) is 5.26% (see bottom-right inset in Fig. 2a).[3] ROC curves for the other populations are available in Supplementary Fig. 3 and have overall a mean AUC of 0.93 and mean false-discovery rate of 6.67% for (see Supplementary Table 1).[4]

1: The first sentence summarises the paragraph.

2: The second sentence provides further information.

3: The third sentence zooms in a particular finding, linked to a figure inset.

4: Finally, the last sentence refers the reader to supplementary information for additional results.

8.4 Discussion

- Use the following structure:

  1¶ Begin by briefly summarising your contribution even if there's a separate conclusion, unless it's for a conference;

  2-3¶ Explain how your work can be generalised to more complex settings, ideally with references to work in Supplementary Information;

  2-3¶ Lastly, back up all the assumptions or hypotheses you made.

- Avoid mentioning 'future work'.

  Avoid the term verbatim, but also paragraphs that apologise for not doing extra work. Instead, explain how your work opens new avenues and

what would be needed for these new directions of research to succeed.

- Avoid mentioning 'limitations' too candidly. The reviewers expect you to offer solutions, justifications, or workarounds.

8.5 Methods

- Methods should be formed of small self-contained units of work.
- Provide simple but replicable descriptions of methods.

  A skilled reader should be able to implement your methodology in their own work.
- In the main text, always briefly introduce each method that will be detailed in Supplementary Information.