

Time for a change in the understanding of what constitutes text plagiarism?

Research Ethics
2015, Vol. 10(4) 187–195
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1747016114552686
rea.sagepub.com


Gert Helgesson

Stockholm Centre for Healthcare Ethics, Karolinska Institutet, Sweden

Abstract

Plagiarism is plaguing research publications in many fields. It is problematic by being misleading about who deserves credit for scientific results, images, text or ideas, by involving scientific fraud (when results are plagiarized) and by distorting meta-analyses. However, different research traditions put different emphasis on the originality of text. Traditional rules regarding correct quotation seem to fit the humanities and many social sciences better than the natural and engineering sciences. This article suggests that we should stop applying a common standard regarding plagiarism to all research fields and instead openly acknowledge that there are differences in what aspects of a paper are important to scientific development in different research areas. More specifically, the article discusses, as a thought experiment, whether the introduction of software supporting text production for research publications in the natural and engineering sciences – thereby further reducing the importance of who created what sentences – would be unacceptable or, quite the reverse, a means to further promote scientific progress. It is concluded that there are no valid principled arguments against introducing such software support for text production in scientific papers, while there are several advantages. Correctly handled, using such software would not involve plagiarism, because it would not be misleading about who deserves credit.

Keywords

plagiarism, research ethics, research traditions, text production

Corresponding author:

Gert Helgesson, Stockholm Centre for Healthcare Ethics, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Tomtebodavägen 18A, Stockholm 171 77, Sweden.

Email: gert.helgesson@ki.se

Introduction

Plagiarism is a well-recognized problem in academia (Titus et al., 2008). Some scientific journals report considerable problems, with up to 30 percent of the papers containing plagiarism (Baždarić et al., 2012; Butler, 2010; Zhang, 2010). There is no universally accepted definition, but according to the core idea in many definitions, plagiarism concerns ‘using someone else’s intellectual product in a way that implies that it is one’s own’ (European Code of Conduct for Research Integrity, 2010; *Longman Dictionary of Contemporary English*, 2013; *Merriam-Webster*, 2013; Office of Research Integrity, 1994; US Federal Policy on Research Misconduct, 2013). A paradigmatic case would be if researcher A includes in his/her paper results, images or text produced by researcher B without informing the readers properly about the origin of this content, thus giving the impression that the material was the intellectual product of A her-/ himself. According to standard citation rules, if text is reused word by word, then this has to be explicitly shown by the use of quotation marks or indentation, in order not to be misleading about who is behind the wording. It is not sufficient in such cases only to provide the correct reference, because this will falsely imply that an idea was borrowed from the cited paper but then formulated in the author’s own words.

Plagiarism is problematic for several reasons: (i) it misleads about the origin and who deserves credit for scientific results, images, text or ideas, and therefore involves deception; (ii) by being misleading about who deserves scientific credit, plagiarism causes an unfair distribution of acknowledged scientific credit (a fair distribution in this context reflects accomplishments) and has potential effects on academic careers and success in obtaining funding; (iii) when *results* are plagiarized, it also involves scientific fraud in the form of fabrication, because by plagiarizing results the authors imply that they have produced those results; (iv) when results are plagiarized and published, it further involves redundant publication because some results are duplicated, with the pretence of presenting original research, which confuses meta-analyses because it gives the impression that more independent research has been carried out than is actually the case (Helgesson and Eriksson, 2014).

Modern information technology makes it easier to plagiarize, because vast numbers of papers, reports, essays and books are easily available in electronic form on the Internet. Whereas transcription by hand was once the standard means of copying others’ work, sometimes requiring tedious work for days on end in university libraries, ‘copy and paste’ is nowadays a swift procedure that can be carried out equally well from the kitchen table. Modern technology also makes it easier to track plagiarism, by the use of software that compares manuscripts to large databases containing vast amounts of published papers, books, etc. (Bechhoefer, 2007; Butler, 2010; Whittle and Murdoch-Eaton, 2008).

Identifying overlapping text void of quotation marks or indentations as plagiarism is, however, to draw too quick a conclusion. The definition given above – ‘using someone else’s intellectual product in a way that implies that it is one’s own’ – shows that there are two ways to avoid plagiarism: either by not using someone else’s product or by not using it in a way implying that it is one’s own. In any research field, there are a number of expressions and sentences frequently used, to describe basic aspects of the methods used, for instance. In many cases these expressions and sentences have been ‘invented’ independently by a large number of authors. When this is the case, such expressions and sentences cannot meaningfully be ascribed to any specific individual – they are not any specific author’s, but a product of many independent minds and therefore a freely shared resource. If an expression or sentence lacks in originality, it cannot be plagiarized, because it does not involve ‘using someone (specific) else’s intellectual product’ (Helgesson and Eriksson, 2014). This means that when plagiarism detection software is used, a judgment is still needed regarding whether or not the overlapping text constitutes an instance of plagiarism.

The other way to avoid plagiarism is by not implying that someone else’s intellectual product is one’s own. This is accomplished by countering the normal implication of including text in one’s paper, namely that it is formulated by the authors themselves. As already mentioned, the standard way to do this is by using quotation marks or indentations for passages borrowed from others. But in principle the implication could also be countered by explicitly stating that the wording is not one’s own. Yet another potential way to counter the implication, I suggest here, would be by a broad agreement to read scientific papers in a way that excludes this assumption.

In this article I will, as a thought experiment, introduce a scenario in which researchers have come to such an agreement. In this scenario scientific authors can freely use text composed by advanced software, based on their needs, to facilitate their writing. The point of the article is to evaluate the implications of such a practice and see what this tells us about plagiarism in general.

Different traditions – different practices

Broadly speaking, there seem to be two different traditions regarding research publications.¹ For the first tradition, all text in a publication is important. For the second tradition, that which is stated in the results section and, to some extent, in the discussion is what really matters, whereas the rest is less important, unless the method itself is the focus of the article. The first tradition dominates in the humanities and many of the social sciences, and the second tradition in much of the natural and engineering sciences.

One could argue that established writing rules relating to plagiarism, applied by universities and publishing houses alike, but also implicit in non-stratified use of

plagiarism identification software, fit the first tradition better: reuse of passages of text without proper recognition of the source comes out as plagiarism regardless of what is reused, whether results or some neat wordings in the introduction or in a footnote.

I suggest that it might be conducive to scientific progress if academia allowed different criteria for plagiarism in different fields, with a focus for the natural and engineering sciences (and perhaps other areas as well) on whether publications are clearly written and not misleading, and if they involve undue acquirement of scientific merit, rather than on whether they contain identical phrasings that are not explicitly quoted. Perhaps different criteria are not needed; it may suffice to recognize that what practices are misleading may vary between different research contexts depending on the expectations of the readers – what is implied by a certain action or practice in one research area may not be implied by the same action or practice in another. For instance, in medicine it seems widely accepted that ‘in medicine we don’t quote’ (as phrased by a doctoral student in a class on research ethics at Karolinska Institutet, Stockholm), by which is meant that researchers provide the reference but do not add any quotation marks, as would be required by traditional writing standards. The general attitude towards reuse of wordings seems to be that it does not matter much whether the cited ideas are rephrased or not, and therefore it could be argued that the ordinary readers in the field are not misled if quotes are used without an explicit recognition of the fact that they are quotes. Doing the same in, say, anthropology, history or philosophy would without hesitation be considered fraudulent.

The thought experiment concerning the use of advanced software to facilitate writing of research papers, which I am now about to introduce, is intended for the tradition that takes text more lightly.

The scenario

Imagine that a number of publishing houses jointly developed software supporting text production in scientific papers. Any researcher ready to start writing with the help of this software would first choose from a fairly large number of drop-down menus to specify area of research, data sample, methods used etc. After specifying the study, the researcher would press a button and the software would produce a few alternatives to choose from regarding introductory text and methods description. It would also provide a suggested structure for the results, discussion and conclusion sections of the article. The researcher would then go on to fill in results, discussion and conclusion using the structure provided for those sections. A standard sentence would be provided in the methods section explaining that text and writing support was provided by the software used, in a vein similar to how statistical programs are acknowledged. No quotation marks or indentations would mark

the software-provided text, so it would not be possible to distinguish between passages written by the authors themselves and passages produced by the software.

In this scenario, not only would such software exist, but its use would be broadly accepted in a number of research areas. In these research areas, it would be agreed that by using the software, and the clarification in the methods section that comes with it, the authors would remove the implication that they are the creators of the phrases and sentences in the article. They may have written quite a bit themselves, but it is agreed that it does not matter whether this is the case or not. What does matter is that the article describes the research correctly and sufficiently.

Assuming, for the sake of argument, that such software could be constructed, let us consider what to say about it. Would it be unacceptable? Or would it be a welcome support in the writing process? Perhaps it would improve the quality of research papers without any considerable downside? Let us look at the arguments in favour of and against introducing such (imagined) software.

Pro-arguments

There are a number of potential advantages with introducing software support of the kind sketched. First, it would solve a practical problem frequently encountered by researchers when writing a paper, namely how to phrase the methods description for yet another paper using the same method. Different solutions are used for this problem today: researchers cut and paste from previous papers (thereby risking accusations of plagiarism or self-plagiarism), leave out most of the methods description by referring to previous papers (thus asking readers to look up another paper to be able to read the present one), or struggle with rephrasing till they run out of alternatives (risking a decrease in comprehensiveness along the way). With the imagined software in place and accepted by the journals and the research community, no one would have to rephrase for the sake of rephrasing. By avoiding this practice, one also avoids the risk of so-called ‘procedure drift’, which refers to a potential risk that research procedures slowly alter owing to a continuous change in published descriptions of the procedure (Jia et al., 2014).

Secondly, it would be an invaluable support for researchers who are good at doing research but have insufficient skills in expressing themselves well in the relevant language. The software would find the best and most established expressions.

Thirdly, with this support, which includes help with structuring the paper and check-lists for required content, the papers can be made more comprehensive and readable for the intended audience. Today there are guidelines for a large number of research fields regarding how results should be reported. Some examples are ARRIVE (guidelines for reporting animal research), the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement, the CONSORT

Statement (on Consolidated Standards of Reporting Trials), the PRISMA Statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and the GRIPS Statement (guidelines for reporting genetic risk prediction studies) (CODEX, 2014). They all provide checklists and, in practice, implicitly provide support on how to phrase the reporting of required content, i.e. authors can borrow ways of constructing sentences by reading the detailed guidelines. Furthermore, some of them explicitly encourage researchers to use publicly available graphic support to better explain the studies; for instance, CONSORT provides flowcharts to illustrate how groups have been allocated to the different arms of clinical trials. If standardization regarding the presentation of tables, graphs, diagrams etc. is laudable because it makes comparisons easier, then by analogy the same should hold for standardization in the writing of scientific papers. Although commented upon in relation to a fairly narrow set of examples, this view seems to be shared by the internationally renowned Committee on Publication Ethics (COPE: Wager, 2011).

Fourthly, by having to spend less time on the very writing of the paper, researchers may instead spend more time thinking about their study and what should be reported about it.

Finally, if research papers are constructed in this way, they can be made machine-readable more easily, which makes them potentially more useful.²

Counter-arguments

There are also potential disadvantages and difficulties with the suggested approach. First, it does not eliminate researchers' language difficulties. There seems to be a considerable risk that researchers with poor skills in the language in which they publish will believe that they have made their research comprehensible without being able to verify that this is the case. Therefore they cannot assume full responsibility for the paper as prescribed in international guidelines (ICMJE, 2013).

Secondly, standardized phrases seem to be a poor tool to handle the need for precision in scientific writing. Although it indeed might be better to explain standard content in the same vein on each occasion, there may be study-specific aspects that cannot be handled by this software. Thirdly, such software arguably paves the way for increased sloppiness, because researchers may become overly confident that the software will take care of all problems connected to getting the message across. Last, but not least, use of the software would lead to plagiarism.

The criticism rightly underlines that use of the imagined software would not eliminate the need to have appropriate language skills. Standardized phrases would not always suffice for a sufficiently detailed description of the study. But they would certainly be of help.

In principle, the language problem could be handled by letting the software run in different languages. The paper could then be translated either by the software or

by authorized translators. If the software were not available in the researchers' native language, the risk remains that they would try to avoid the cost for translation by using the software as best they could. Also, considering the innate problems with automatic translation, the texts would probably still have to be language-reviewed if translated by the software.

The risk of sloppiness must also be admitted, but it is always present with or without the imagined software and has to be countered by creating good research environments. With less time needed for writing, more time could be spent on considering how results are presented and on discussing their implications.

What about plagiarism? If plagiarism is understood as breaking traditional rules for appropriate referencing, then use of the software as described above would involve plagiarism. This is so because which passages are generated by the software and which are produced by the researchers themselves would not be clearly distinguishable. However, if plagiarism instead were understood as in the suggested definition ('using someone else's intellectual product *in a way that implies that it is one's own*'), then it is reasonable to argue that if this had become well established and if it were clearly stated that the (imagined) software had been used in text production, then the authors would not be implying that they had written the text themselves, even if they did not use quotation marks, while they would be implying that they had carried out the research presented.

Conclusion

The presented thought experiment, regarding a systematic use of software support for text production in scientific papers, shows that there are no principled arguments against such an approach to scientific text production. If a standard explanation and a proper reference to the software were given, and its use was widely accepted within the research field, then this would remove the implication that the authors of the paper had produced the text themselves. Thus, use of such software would not be misleading and would not involve plagiarism.

The more general implication of the discussion is that if certain text production serves the purpose of signalling what kind of research it is by giving a brief, correct description of the context for the research or the chosen methods, then it is preferable that these passages are standardized rather than carrying the distinctive marks of the individual authors. For such text it would be better if the research community could openly agree to disregard where these phrases and sentences originate, rather than continuing to scrutinizing them in order to identify plagiarism, interpreted as overlapping text without quotation marks.

Regardless of whether or not the fictitious software ever comes into existence, researchers have to take full responsibility for what they submit to journals and make sure that sloppiness and carelessness do not get in the way of a correct and transparent presentation of their research.

Acknowledgements

I would like to thank Stefan Eriksson for many inspiring discussions on plagiarism and Tomas Månsson for valuable comments on a previous version of this article.

Declaration of conflicting interest

The author declares that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. This is a conjecture, but not a particularly bold one. However, what matters for the argumentation is that there are research fields where the origin of wordings is perceived as of minor importance.
2. This was pointed out to me by an anonymous reviewer.

References

- Baždarić K, Bilić-Zulle L, Brumini G and Petrovečki M (2012) Prevalence of plagiarism in recent submissions to the Croatian Medical Journal. *Science and Engineering Ethics* 18(2): 223–239.
- Bechhoefer J (2007) Plagiarism: Text-matching program offers an answer. *Nature* 449(7163): 658.
- Butler D (2010) Journals step up plagiarism policing. *Nature* 466(7303): 167.
- CODEX (2014) Available at: www.codex.vr.se – a website dedicated to research ethics – (accessed 15 May 2014).
- European Code of Conduct for Research Integrity (2010) Available at: http://www.allea.org/Content/ALLEA/Scientific%20Integrity/Code_Conduct_ResearchIntegrity.pdf (accessed 20 December 2013).
- Helgesson G and Eriksson S (2014) Plagiarism in research. *Medicine, Health Care and Philosophy*. Epub ahead of print, 4 July, doi:10.1007/s11019-014-9583-8.
- International Committee of Medical Journal Editors (2013) ICMJE Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. Available at: http://www.icmje.org/new_recommendations.html (accessed 15 May 2014).
- Jia X, Tan X and Zhang Y (2014) Replication of the methods section in biosciences papers: Is it plagiarism? *Scientometrics* 98: 337–345.
- Longman Dictionary of Contemporary English*. Available at: <http://www.ldoceonline.com/dictionary/plagiarism> (accessed 20 December 2013).
- Merriam-Webster*. Available at: <http://www.merriam-webster.com/dictionary/plagiarism> (accessed 20 December 2013).
- Office of Research Integrity (1994) ORI policy on plagiarism. *ORI Newsletter* 3(1).
- Titus SL, Wells JA and Rhoades LJ (2008) Repairing research integrity. *Nature* 453(7198): 980–982.
- US Federal Policy on Research Misconduct. Available at: <http://www.aps.org/policy/statements/federalpolicy.cfm> (accessed 20 December 2013).

- Wager L (2011) *How should editors respond to plagiarism?* COPE discussion paper, 26 April. Available at: <http://publicationethics.org/files/Discussion%20document.pdf>.
- Whittle SR and Murdoch-Eaton DG (2008) Learning about plagiarism using Turnitin detection software. *Medical Education* 42(5): 528.
- Zhang Y (2010) Chinese journal finds 31% of submissions plagiarized. *Nature* 467: 153.