Defining and responding to plagiarism

Elizabeth WAGER
Sideview

ABSTRACT. *A clear definition of plagiarism and the ability to classify it into more or less serious forms would help editors and publishers to devise policies to handle this problem. Text-matching software is a useful tool for measuring the extent of text copying but it cannot detect plagiarized tables or figures, plagiarism of ideas, or plagiarism in translation. Furthermore, a working definition of plagiarism in relation to research reports needs to take into account factors such as the originality of the copied material, its position in the report, the adequacy of referencing, and the intention of the authors as well as the extent of the copying. This article considers all these factors and proposes possible definitions of major and minor plagiarism in relation to scholarly publications which might be used as the basis for anti-plagiarism policies in conjunction with resources such as the COPE flowcharts.*

Introduction

Plagiarism has caused problems for editors and publishers for centuries. The origin of the word may be traced back to the Roman poet Martial (*c.* AD 80), who claimed that another poet recited 'my books to the crowd as if none other than your own'. The term was probably first used in English, in its current sense of literary theft, sometime in the 15th century.[1] In the 21st century, awareness of, and interest in, plagiarism has been fuelled by the availability of text-matching software which can reveal the use of 'copy–paste' by authors. However, while software (such as iThenticate, eTBlast, or use of search engines) makes it easy to detect text matches, not all such matches are necessarily plagiarism. Text matches may occur for legitimate reasons such as: co-publication of guidelines (agreed by several journals), republication with the agreement of original and new publishers to make documents available to new audiences (as happened, for example, with this article), online availability of conference abstracts resembling later journal publications, or the existence of several versions of a document (e.g. in preprint servers or institutional repositories as well as journal websites). Reuse of short word strings may be accidental or unavoidable, and exact repetition of technical descriptions (e.g. in a methods section) may be helpful to ensure accuracy. Therefore, editors and publishers need workable definitions of plagiarism and probably a more detailed taxonomy to distinguish the different forms of plagiarism than that provided by dictionary definitions.

Once they have defined plagiarism, and armed with sophisticated tools for detecting it, editors need to decide how to handle it.

Background/existing guidelines

The COPE flowcharts[2] recognize that an editor's response to plagiarism should depend on

Elizabeth Wager

the type and extent of the copying. They suggest different responses to 'Clear plagiarism' (described as 'unattributed use of large portions of text and/or data, presented as if they were by the plagiarist') and 'Minor copying of short phrases only' with 'no misattribution of data' (giving an example of copying 'in [the] discussion of [a] research paper from [a] non-native language speaker'). The flowcharts also distinguish plagiarism (i.e. copying from others) from redundancy or 'self-plagiarism' (i.e. copying from one's own work). The flowcharts also suggest that the editor's response might vary according to the seniority of the author (with editors simply writing an educational letter to very junior researchers but considering informing the institution of more senior authors) as well as whether the authors are writing in their native language.

Types of plagiarism

Any original creation may be plagiarized. Although most discussions focus on text (and this type of copying is usually the only kind that can be detected using software) it is important to recognize that ideas, images, creative works (e.g. musical compositions or choreographies), and data can be plagiarized.

The following factors may be helpful in distinguishing types of plagiarism (see Table 1):

- extent
- originality of copied material
- position/context

- referencing/attribution
- intention
- author seniority
- language.

Extent

The most blatant forms of plagiarism involve the copying of entire papers or chapters which are republished as the work of the plagiarist. Such cases usually involve not only plagiarism but also breach of copyright. Whole articles or chapters may also be plagiarized in translation. The COPE retraction guidelines recommend that such articles should be retracted, and the flowcharts on plagiarism suggest that editors should consider contacting the author's institution in such cases.[3] However, the COPE retraction guidelines state that:

> [I]f only a small section of an article (e.g. a few sentences in the discussion) is plagiarised, editors should consider whether readers (and the plagiarized author) would be best served by a correction (which could note the fact that text was used without appropriate acknowledgement) rather than retracting the entire article which may contain sound, original data in other parts.

Scholarly works often summarize the work of other researchers. It may be difficult to draw a line between legitimate (and accurate) representation of other studies and copying original material. Researchers may also feel that little harm is done if they use similar language to another publication so long as the source is

it is important to recognize that ideas, images, creative works (e.g. musical compositions or choreographies), and data can be plagiarized

Table 1. Features of different types of plagiarism

Feature	Least severe type ————————————————————▶				Most severe type
Extent	A few words	A few sentences	Whole paragraph	Several paragraphs	Whole paper
Originality of copied material	Widely used phrase/idea		Phrase/idea used by a small number of authors		Original phrase/idea
Position/context/type of material	Standard method		Describing another worker's findings		Data/findings
Referencing/attribution	Source fully and clearly referenced		Source partially/inaccurately referenced		Unreferenced
Intention	No intention to deceive				Intention to deceive

Table 2. Examples of language of low originality used in reports of medical research

Phrase	Hits for exact phrase (January 2011)	
	Google	Google Scholar[a]
$P < 0.05$ was considered statistically significant	588,000	70,600
performed according to the Declaration of Helsinki	410,000	1,860
double-blind, double-dummy, placebo-controlled	56,800	882
numbered, opaque, sealed envelopes	12,200	912
randomized in a 1:1 ratio	8,510	1,020
computer-generated random number list	5,120	354

[a]Google Scholar searches for academic publications only.

properly cited. If the original authors summarized their findings clearly and succinctly it could be argued that little is gained by forcing other authors to paraphrase this. However, others will argue that any verbatim copying should be indicated by using quotation marks, otherwise they would consider it to be plagiarism.

Most text-matching software detects strings of several words, since duplication of just a few words can occur by chance. However, academic papers and reports may contain technical language that involves standard phrases that are longer than the strings used by software. For example, a Google search for the phrase 'smokers with chronic obstructive pulmonary disease' produces >58,000 results, suggesting that this is a widely used phrase, but such a six-word string may also trigger a match on text-matching software. Therefore extent alone cannot be taken as a benchmark. Table 2 shows the frequency of some phrases commonly seen in medical publications which are so widely used their use would not be considered as plagiarism but which might count as matches when using text-matching software.

extent alone cannot be taken as a benchmark

Originality of copied material

Originality needs to be considered in conjunction with extent. The example given above indicates the difference between a standard, technical phrase (such as 'smokers with chronic obstructive pulmonary disease') and original usage (such as Winston Churchill's 'the end of the beginning' or Shakespeare's 'the winter of our discontent' – both of which contain fewer than six words and would therefore probably not be detected by text-matching software yet are usually considered sufficiently

original to be attributed to the creator). While publishers of poems and song lyrics tend to guard their copyright fiercely, and permission is required to quote even a single line, technical publications may contain descriptions of standard techniques which are described in similar or identical ways because the language is not considered original. Therefore, the originality of the copied material should be considered as well as the extent.

Position/context

Certain sections of research reports may be more likely to include non-original material. In particular, the methods section may describe widely used techniques. The use of standardized descriptions of public data sources, proprietary techniques, questionnaires, or equipment may even be regarded as helpful to ensure accuracy and consistency. For example, analysis of the UK General Practice Research Database has resulted in over 750 publications. All these publications probably include a description of the database and these are likely to use similar language. Similarly, it may be better if the original description of an assay provided by a company or the supplying laboratory is copied rather than reworded by each user, since the original wording may be the most accurate. Therefore editors may view text similarity in methods sections as unavoidable or even desirable and therefore treat it differently from similarity occurring in other parts of a paper.

The editor of a mathematics journal has noted:

> [S]tatements of the mean value theorem from calculus book to calculus book are virtually identical; there's really only way to state

Schur's Lemma. Probably, automated software would detect these instances as plagiarism. They're not, of course. Sometimes, there's really only one way to define something or give the 'usual examples'. (Lance Small, personal communication)

The type of publication may also affect judgements about the acceptability of text similarity. While research reports may describe standard methods, editorials may be expected to represent the author's opinion and original views on a topic and it would therefore be considered inappropriate to use the same words as another author except in direct and properly attributed quotations. Similarly, review articles, and the discussion sections of research papers, are expected to provide an original synthesis of, and commentary on, previously published work. Therefore, apart from quotations, the words may be expected to be the author's own.

However, editors may also consider the consequences of the copying and its potential to mislead readers. In this respect, copying a few sentences from the discussion section of another researcher's paper may be considered less harmful, and less deceitful, than the theft of data (which may constitute not only plagiarism but also data fabrication since the work being reported was not actually done by the copier). Thus, if a submitted article that appears to describe legitimate, original research, includes some sentences taken from the discussion of another author's paper on a related topic, the editor may simply ask the author to indicate that these are direct quotations, or to paraphrase the copied text, before publication. If the copying is discovered after publication, the editor may suggest that it can be rectified by a correction rather than a retraction and may not feel that the author's institution should be informed.

When using software to detect text similarity, editors should not forget that reference sections will contain large amounts of copied text in the titles of cited articles. Some software systems, such as iThenticate/CrossCheck, allow these sections of the paper to be excluded from the search, together with any text enclosed in quotation marks.

If systematic reviews or databases are updated, the original authors may be replaced by others. An updated review or database will, naturally, contain large sections from the previous versions and this may appear to be plagiarism if the authors have changed (since automatic systems will not recognize acknowledgements to previous versions).

Referencing/attribution

Academic publications are expected to reference other works and may also quote from them. Inexperienced or poorly trained authors may mistakenly believe that so long as another work has been cited, parts of it can be reproduced in their own work. While copying parts of cited work is probably not intended to deceive the reader in the same way as copying unattributed material, the practice is generally considered to be poor scholarship and inappropriate for an academic journal. Editors may have a role in educating authors if they discover this type of copying, especially if it is detected before publication.

Intention

Intention to deceive is often considered a factor distinguishing misconduct from careless work or honest error. However, it is usually impossible to prove intent and therefore may be less useful in practice than in theory. Extreme forms of plagiarism, such as copying an entire paper and submitting it under a different author's name to another journal, can only be deliberate. Editors must use their own judgement to determine whether authors' explanations for less extreme forms of copying are plausible or could have occurred through honest error. When confronted with identical text, authors may counter with explanations such as having a photographic memory or inadvertently copying notes or preliminary material into a publication. When a senior researcher at Stanford University was found to have incorporated large chunks of text from a well-known textbook into a chapter he had prepared for another book, he told an inquiry that 'when he cut-and-pasted the material into his manuscript, he added handwritten notations detailing where the text came from. These notations were supposed to have been printed in the body of his chapter'. Nevertheless he was found guilty of 'grossly negligent scholarship' and resigned as chairman of the Department of Medicine.[4]

inexperienced or poorly trained authors may mistakenly believe that so long as another work has been cited, parts of it can be reproduced in their own work

Authors who admit intentional copying may nevertheless insist that this is acceptable in their discipline or culture and that, rather than representing academic theft or laziness, it is, in fact, a form of flattery or 'homage' to the original author. They may also suggest that quotation marks are unnecessary because specialist readers (for whom they are writing) will immediately recognize the quotations and be aware of their source. However, editors and readers may find these arguments unconvincing.

Author seniority

Since editors may believe that some forms of plagiarism result from poor mentorship or supervision rather than intentional misconduct, their response may vary according to the seniority of the authors involved. Editors may apply different sanctions to junior authors who they believe genuinely did not realize they were doing something inappropriate from those applied to experienced researchers who are expected to know better. Thus, an editor may respond to the copying of a paragraph from a cited paper by asking a junior author to paraphrase (if detected before publication) or issue a correction (if detected after publication). However, for a similar degree of copying by a senior author, the same editor might reject or retract a submission and consider informing the author's institution.

Informing an author's institution is generally considered to be a relatively serious action to take, since it may have grave consequences for the researcher concerned. Editors therefore tend to be reluctant to inform institutions except in serious cases of misconduct and when they feel they have well-founded suspicions of wrong-doing. However, if contacting an institution is viewed, not as a potential punishment for the author, but as an attempt to engage the institution in dialogue and work together to prevent future problems, one might argue that editors should contact institutions more often and definitely in cases where they feel junior researchers have received inadequate training or guidance, since this is something the institution may be able to remedy. Before contacting an institution it is advisable to ask the author(s) for an explanation even in the face of strong evidence of apparently blatant plagiarism. This is recommended by the COPE flowcharts based on experience that apparently 'cut-and-dried' cases of misconduct often turn out to be more complex than they appear on the surface.

If an editor detects copying in a manuscript that is going to be rejected, then contacting a head of department or dean might prevent the authors from simply submitting the manuscript, unchanged, to another journal.

Language

Text-matching software will only detect text copying in the same language. However, republication of an unattributed translation of another person's work is also plagiarism, although it is harder to detect and may be harder to prove unless extensive.

Just as editors' responses may depend on the authors' seniority, they may also depend on whether authors are writing in their native language since editors recognize the difficulties that non-native writers face in correctly paraphrasing other authors' work. In some cases, researchers may actually have been encouraged, when learning a language, to adapt sentences and 'borrow' structures from published works. This may result in so-called 'patch' (or 'patchwork') writing.[5] This form of copying will only be detected by sensitive text-matching systems and those that employ a degree of 'fuzzy' matching, since authors are likely to have changed some words in adapting the sentence for their own use. Authors who use this technique usually copy from a wide range of sources, often with individual sentences coming from different publications. This may result in a high total similarity 'score' for the article from an anti-plagiarism detection system such as CrossCheck, but the matched text will be found to come from many sources, and each copied section will be short (with few or no substantial chunks of copied text). However, few, if any, of the sources of the copied text are likely to be cited in the publication, since they may be on unrelated topics.

Some editors may see little harm in authors who describe their own methods and findings accurately, but use sentence structures taken from other publications. Others may regard this as a sign of poor scholarship or a form of minor plagiarism. The acceptability of 'patch' writing probably depends on the originality of the writing being copied. While

republication of an unattributed translation of another person's work is also plagiarism, although it is harder to detect

it may be entirely unacceptable for works of creative fiction, it may be considered acceptable when describing widely used methods which, as already mentioned, may require a degree of standardized text to ensure accuracy. If the copied structures are clear and grammatically correct, some editors may even feel that this method of writing will benefit readers and journals, since methods will be accurately described and the manuscript will require less copyediting to correct grammatical mistakes.

However, others may have concerns that authors will be tempted to copy inappropriate phrases that do not correctly describe their own research, especially if they do not completely understand the phrases being copied. To misrepresent research methods (e.g. by stating that a study was prospective or randomized when, in fact, it was not) is generally considered a serious form of misconduct. Editors may therefore be concerned that, if they tolerate 'patch' writing, such misrepresentation may go undetected.

screening incurs costs, and therefore editors and publishers need to decide the best ways of employing it

Detecting and responding to plagiarism

The original COPE guidelines on good publication practice (published in 1999) noted that 'plagiarism ranges from the unreferenced use of others' published ideas … to submission under "new" authorship of a complete paper, sometimes in a different language'.[6] However, these guidelines did not describe what editors should do if they encountered these different forms. The guidelines did offer general guidance on the sanctions that editors might take against authors (see text box).

The COPE flowcharts on plagiarism (published in 2006) recommend different responses for 'clear plagiarism' and 'minor copying' but provide only rather general indications of how editors might distinguish these two phenomena.[2]

Screening for plagiarism

The availability of powerful tools such as CrossCheck makes it possible to screen submissions for matching text and some journals are now doing this routinely. However, screening incurs costs (in the form of charges for using the tools, and in terms of editorial time), and therefore editors and publishers need to

---

Possible responses to misconduct

(From COPE, *Guidelines on Good Publication Practice*, 1999)

The following [sanctions] are ranked in approximate order of severity:

(1) A letter of explanation (and education) to the authors, where there appears to be a genuine misunderstanding of principles.

(2) A letter of reprimand and warning as to future conduct.

(3) A formal letter to the relevant head of institution or funding body.

(4) Publication of a notice of redundant publication or plagiarism.

(5) An editorial giving full details of the misconduct.

(6) Refusal to accept future submissions from the individual, unit, or institution responsible for the misconduct, for a stated period.

(7) Formal withdrawal or retraction of the paper from the scientific literature, informing other editors and the indexing authorities.

(8) Reporting the case to the General Medical Council, or other such authority or organisation which can investigate and act with due process.

---

decide the best ways of employing it. The options include:

- screening all manuscripts on receipt
- screening manuscripts that are sent out for external peer review
- screening manuscripts that are provisionally accepted
- screening a random sample of manuscripts
- using the software only in cases when plagiarism is suspected.

Anecdotal reports suggest that, when CrossCheck was first available, many journals started by screening only accepted manuscripts but later switched to screening all submissions because of the frequency of problems they discovered (since they did not wish to send peer-reviewers manuscripts that were later found to raise concerns about plagiarism or redundancy). The Editor of *Anesthesia & Analgesia* noted in an editorial that 'I have screened

every submitted manuscript for many months. Approximately 1 of every 10 submissions has had unacceptable amounts of text taken verbatim and without attribution from another source.' Elsevier has announced that, as from 2014, it will use CrossCheck on all submissions and has integrated the text-matching software into its editorial system.[7,8]

Defining plagiarism

Editors need to decide how to interpret and respond to findings of text similarity. It is important that authors receive fair and consistent treatment from journals, but devising a detailed policy on responses to plagiarism is difficult given the many forms that plagiarism can take. Because text-matching software has only become available relatively recently, cases of plagiarism are very likely to be uncovered in back issues of the journal. Editors therefore need a clear policy for responding to plagiarism in material published recently and further back in the past.

Clear-cut cases of serious plagiarism (e.g. whole articles or large sections of text) may warrant retractions. Since the general concept of plagiarism is not new, and large-scale plagiarism has been identified as a serious form of misconduct for decades, most editors would agree that this is the correct course of action. However, identification of 'patch writing' or 'micro-plagiarism' has only become possible with the availability of specialized software. Some editors may therefore feel uncomfortable about applying sanctions to authors retrospectively. One solution to this problem would be to announce an amnesty for older publications (i.e. an agreement that the journal will not take action if minor plagiarism is found in previous issues) but warning authors that text similarity in future submissions will not be tolerated.

To devise workable policies for both submitted and published articles, editors need to consider the thresholds for deciding when to:

- educate authors and ask them to rewrite copied text
- reject an article
- issue a correction (for a published article)
- issue a retraction (for a published article)
- inform an author's institution.

Considering the COPE flowcharts, maybe they should provide more guidance to help editors distinguish major from minor plagiarism. One possibility would be to produce definitions based on the characteristics described above.

For example, *major plagiarism* could be defined as any case involving:

- unattributed copying of another person's data/findings, *or*
- resubmission of an entire publication under another author's name (either in the original language or in translation), *or*
- verbatim copying of >100 words of original material in the absence of any citation to the source material, *or*
- unattributed use of original, published academic work, such as the structure, argument, or hypothesis/idea of another person or group where this is a major part of the new publication and there is evidence that it was not developed independently.

Minor plagiarism could be defined as:

- verbatim copying of <100 words without indicating that these are a direct quotation from an original work (whether or not the source is cited), unless the text is accepted as widely used or standardized (e.g. the description of a standard technique)
- close copying (not quite verbatim, but changed only slightly from the original) of significant sections (e.g. >100 words) from another work (whether or not that work is cited).

Use of images without acknowledgement of the source could be defined as:

- republication of an image (photograph, diagram, drawing, etc.) generated by another person without acknowledging the source.

Journal responses could then be matched to these. For example:

- *Minor plagiarism in a submitted article* – write to author and request reworking or (if article is being rejected) point out that minor plagiarism has been detected and advising the authors that this should be corrected before resubmission.
- *Minor plagiarism in a published article* – contact author and discuss findings, and issue a correction and apology. It is important to ensure that corrections are clearly indicated

clear-cut cases of serious plagiarism (e.g. whole articles or large sections of text) may warrant retractions

in all versions of the publication (i.e. both online and print) and, where possible, via databases and indexing systems – the use of systems such as CrossMark offers possibilities for alerting readers to corrections even if they are using a downloaded PDF.

- *Major plagiarism in a submitted article* – present findings to all authors and ask them to respond; ask the authors if all or only some of them are responsible for the plagiarized sections, decide if any authors were unaware of the plagiarism and, if so, whether they are in any way responsible for the behaviour of the other authors (e.g. in a supervisory capacity); explain that plagiarism is unacceptable and that you plan to inform their institution; contact the institutions of authors you consider were directly involved with, or should take responsibility for, the plagiarism

- *Major plagiarism in a published article* – as for submitted article, then retract article.

- *Use of images without acknowledgement of the source* – if the image contains data from another person's research (e.g. a graph), and this is shown as if it were the work of the copyist, this should be treated as data copying (i.e. major plagiarism). For images that do not contain original data (e.g. diagrams showing processes, maps, illustrative photographs), the author of a submitted paper should be told to seek permission for republication from the copyright holder, remove images for which permission is not granted, and insert appropriate acknowledgements for images for which permission has been granted; if such images have already been republished, the editor should contact the author and issue a correction giving the appropriate acknowledgements.

Conclusions

Editors and publishers should consider their policies on detecting and handling plagiarism. To be helpful, and to be applied consistently, such policies need a workable definition of plagiarism. This article offers some possible suggestions. Text-matching software coupled with extensive databases can create powerful tools such as CrossCheck but publishers need to decide how best to use these. Two studies presented at the 2013 Peer Review Congress assessed the amount of time needed to screen articles.[9,10] Both concluded that screening was feasible and useful, but also noted that it required an investment of time which, for journals receiving large numbers of submissions, could be substantial.

References

1. Bailey, J. The world's first plagiarism case. http://www.plagiarismtoday.com/2011/10/04/the-world%E2%80%99s-first-plagiarism-case/ (accessed 2 October 2013).
2. Committee on Publication Ethics. Flowcharts 'What to do if you suspect plagiarism' http://publicationethics.org/files/u2/02A_Plagiarism_Submitted.pdf and http://publicationethics.org/files/u2/02B_Plagiarism_Published.pdf. 2008 (accessed 2 October 2013).
3. Committee on Publication Ethics. Retraction guidelines. http://publicationethics.org/files/retraction%20guidelines.pdf (accessed 2 October 2013).
4. Norman, C. 1984. Stanford investigates plagiarism charge. *Science*, 224: 35–36.
5. Kerans, M.E. and de Jager, M. 2010. Handling plagiarism at the manuscript editor's desk. *European Science Editing*, 36: 62–66.
6. Committee on Publication Ethics. *Guidelines on Good Publication Practice*