INVITED EDITORIAL
doi: 10.2306/scienceasia1513-1874.2018.44S.007

ScienceAsia 44S (2018): **7–10**

Automated checking and editing of LaTeX manuscripts

Michael A. Allen

Physics Department, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400 Thailand

e-mail: maa5652@gmail.com

ABSTRACT: After outlining the post-referee manuscript processing scheme *ScienceAsia* uses, a brief description is given of various pieces of software developed by the author to perform automated checks on and edits of LaTeX files.

INTRODUCTION

LaTeX (pronounced "lay-teck")[1] is the document preparation system of choice for typesetting books or journals, and particularly those with a high mathematics content; it is favoured by most mathematicians, physicists, and the publishers of their work[2] and is more efficient in the production of these types of document[3]. This year marks the 40th anniversary of the initial release of TeX ("teck")[4], the system on which LaTeX is built, and the 80th year of TeX's polymath originator, Donald E. Knuth. It is also 10 years since *ScienceAsia* switched to typesetting entire issues via LaTeX (while continuing to permit authors to submit non-mathematical manuscripts prepared using MSWord). The switch was made partly because LaTeX, although free, produces beautiful, professional-looking output, but mainly because it has enabled much of the editorial work to be done automatically. The automation of routine tasks where possible has proved to be vital since the journal has for long periods suffered from the lack of a full-time staff member as a result of funding problems.

It has been commented in a similar column to this that a growing interest in publication statistics has led to some authors appearing to take a greater pride in the quantity rather than the quality of their output[5]. The rising number of poorly written papers has resulted in editors being overworked and it is now not unusual to see grammatical errors in even the titles of articles from journals of respected publishers such as *Elsevier*. A fully automated scheme for correcting text would require sophisticated artificial intelligence. However, there are many types of simple mistake that are easy to detect, and in most cases correct, automatically, leaving editors to concentrate on finding the more subtle errors.

This article is intended to raise awareness among developers of editing software and publishers that automated checking and editing is possible and desirable. After giving a brief overview of LaTeX and how *ScienceAsia* handles manuscripts after they have been refereed, we describe the operation and rationale behind the two most important programs the journal uses to process LaTeX input files: `ckms` ("check-em-ess") and fixTeX ("fix-teck"). Both are C programs developed by the author for the Linux operating system.

LaTeX

The philosophy behind LaTeX, as opposed to WYSIWYG word processors like MSWord, is that the ordinary user should not have to deal much with the format of the document (which will be taken care of by the publisher) and so they can pay more attention to the most important part, the content. LaTeX is a simple programming language although this aspect is only apparent when material other than ordinary continuous text is required. A typical input file (which LaTeX reads to create the final document) consists mostly of the desired prose interspersed with 'commands' used, for example, to obtain special symbols, create or refer to equations or tables, or include graphics.

The input (`.tex`) files that LaTeX acts upon are text (ASCII) files. This is the essential advantage of LaTeX from the point of view of carrying out automated checks and edits since such files are easily read or written to by programs without decoding or encoding, unlike, e.g., `.doc` or `.rtf` files.

A `.tex` file consists of commands, comments, and ordinary text. Commands generally start with a backslash ('\') character and may have arguments which are enclosed by curly (and sometimes square) brackets. Commands correspond to the subroutines or functions of a typical programming language. An important class of commands are those that begin and end a LaTeX environment. These are used to delineate objects such as figures, tables, and mathematical expressions. The simplest commands have no arguments. For example, `\alpha` gives the Greek

letter α when used inside a maths environment.

Anything after a % (which is not immediately preceded by an odd number of backslashes) on a line is ignored by LaTeX and so such material is classed as a comment. Comments are used, for example, in the template `.tex` file to tell authors where to put various pieces of information such as names and affiliations.

Anything which is not a command or comment is transferred letter by letter to the output. Blank lines are used to indicate a new paragraph. For further details on LaTeX and how the journal suggests that it is used see www.scienceasia.org/scias_latex.pdf.

Class files

A complete `.tex` file for processing by LaTeX starts by specifying the document class (`.cls`) file. This tells LaTeX how to handle the standard commands such as `\author{ }` and `\title{ }`, and also allows the definition of non-standard commands tailored to meet the specifications of the document being prepared. *ScienceAsia* has its own two class files: one for individual articles (`scias.cls`) and one for producing the entire issue. The use of such class files means that as much as possible regarding formatting and numbering can be automated which makes life easy for the authors and editors alike.

POST-REFEREE MANUSCRIPT PROCESSING

After the referees are satisfied with the manuscript and the authors have sent the manuscript source files, it undergoes what *ScienceAsia* refers to as a post-referee review whereby an editor makes further checks on the manuscript and usually requests further changes from the authors with the assistance of the program `ckms`. When the manuscript is prepared using LaTeX `ckms` performs many of these checks automatically.

On completion of the requested changes, in the case of manuscripts prepared using MSWord, the `.doc` file of the manuscript is converted into a `.tex` file using the free software package Writer2LaTeX. The `.tex` file so produced tends to be 'messy' in the sense that it is full of redundant LaTeX commands which make it more difficult to edit. The file is therefore tidied using the program fixTeX run in 'safe mode' which means that none of the manuscript text is changed. Then fixTeX is applied to all manuscripts in the normal mode in combination with the free program `meld` which highlights any changes made to the manuscript by fixTeX. If a particular change is not desired, it can then be reversed with the click of a mouse button.

Additional programs are run on the `.tex` file to order and, where possible, adjust the formatting of the references before the `.tex` file is edited by hand. `ckms` is run on the file again to check for errors introduced accidentally during manual editing. A spell-checking program (which is instructed to ignore LaTeX commands and in some cases their arguments) is also run on the file prior to producing the proof version sent to the authors.

Once all the manuscripts for one journal issue are ready, LaTeX is run on a file giving the order in which the articles appear and creates the entire issue apart from the cover which is obtained as a separate PDF file, again via LaTeX. Finally, a program is run to extract from the various files involved the information needed for the database used by the journal website for the online issue.

AUTOMATED CHECKING: `ckms`

Operation

LaTeX is first run on the `.tex` file as, in addition to the PDF file, it produces a `.log` file listing certain types of error which is later read by `ckms`. `ckms` is then run from the command line. By default, it opens a graphical user interface with a text window listing errors it finds automatically (Fig. 1). This text can be edited by typing or have set responses added by pressing labelled buttons. Once this is done, the contents of the text window is emailed to the author.

Rationale

The philosophy of using `ckms` and fixTeX sequentially is that, in general, the authors are not asked by `ckms` to correct anything that fixTeX can already safely take care of later. As fixTeX and the reference fixing software develops, the number of types of request that `ckms` can issue will diminish.

Automated checks on label and ref commands

All figures, tables, numbered equations, and theorem-like structures have a number associated with them by LaTeX. This number can be referred to using the label and ref commands and that way if such a numbered entity is added or removed, the renumbering is redone automatically by LaTeX. The *ScienceAsia* class file is set up so that such references are hyperlinked as well. The journal therefore insists that authors use these commands appropriately and `ckms` warns if this is not the case.

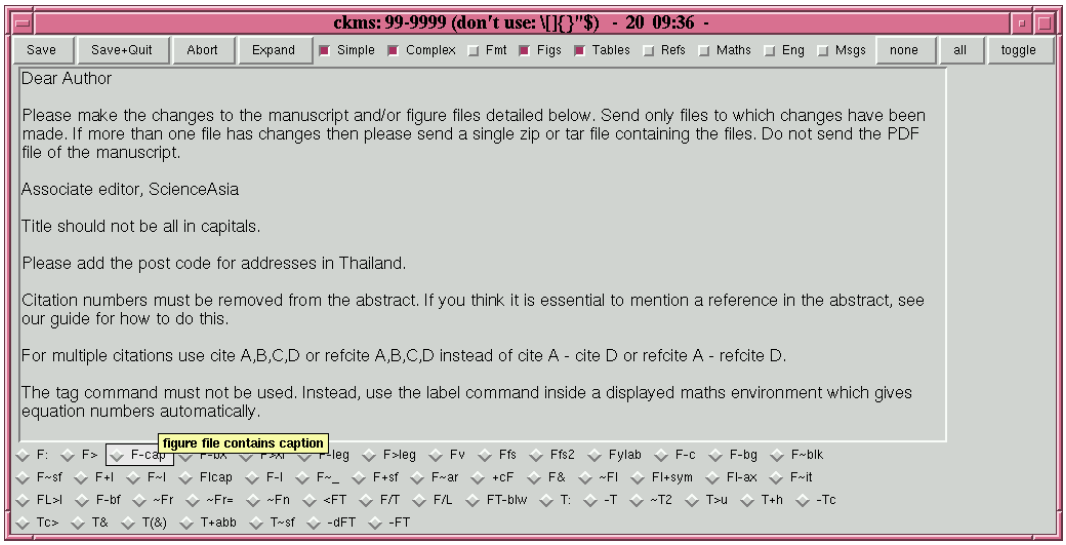

Fig. 1 The `ckms` graphical user interface showing some errors in a manuscript `.tex` file detected automatically.

Automated checks that appropriate LaTeX environments are being used

The *ScienceAsia* class file has pre-defined LaTeX environments for theorem-like structures and proofs which result in uniformity of appearance across articles. `ckms` checks that these are being used where appropriate.

Automated checks on commands affecting the default format of an article

Authors cannot be allowed to redefine the standard LaTeX commands as this could affect the articles in the issue that follow it when the PDF of the entire issue is prepared for the printing house. `ckms` therefore flags any such attempts.

`ckms` also checks for the inclusion of packages that are incompatible with/and or affect the layout the journal uses and checks that the author is following the journal guidelines regarding such things as appendices, footnotes, or bullet points.

Automated checks on formatting equations

If equations are too long and start to enter the right margin then this is mentioned in the `.log` file. If this is to a significant extent, `ckms` will ask the author to adjust the equations (normally by splitting them onto more than one line) to avoid this.

Automated checks on citations

`ckms` flags any entry in the list of references that is not referred to in the text. This can occur if the author uses a dash to specify a range of reference numbers (rather than list them explicitly within the cite command and let LaTeX add a dash if appropriate). `ckms` warns if this occurs. Conversely, any references made to items that do not occur in the list will generate an error message (which is stored in the `.log` file) when LaTeX is run and `ckms` will notify the user about these too.

With systems like BibTeX for LaTeX and EndNote for MSWord users, it is straightforward to put the references in the correct format. If authors are unwilling to use these systems, the journal insists that the authors correct the format manually. Some reference format problems which cannot be fixed automatically are detected by `ckms`.

Automated checks on spelling

Common misspellings (such as 'seperate') are taken care of by fixTeX, as are non-international English variations in spelling. Other misspellings of English words which are not proper nouns are picked up with the help of spell-checking software in the final stages of editing.

Proper nouns in manuscripts unlikely to already be in the spell-check database are mostly names of cited authors and the verification of their spelling is time-consuming to do manually. However, proper nouns are easily identified when not at the start of a sentence since they always start with a capital letter. `ckms` warns if two proper nouns differ by one letter or by the interchange of two of the letters (excluding the first two).

In the case of foreign words, accents such as umlauts can only ever occur on certain letters. `ckms` locates any occurrences of accents that break these rules.

Automated checks on English

Although some grammatical errors are corrected by fixTEX, and many can be corrected by the editors, there are a significant fraction which can only be corrected by someone who has sufficient knowledge of the field. It is therefore far preferable that the authors get the grammatical errors ironed out with the help of an English expert. `ckms` looks for words or phrases which are always wrong (such as 'in term of' or 'informations') and then warns that the manuscript contains grammatical errors. This is done without specifying the examples it has found as experience shows that some authors will then only correct those points.

It was recently reported in a light-hearted article that authors have got away with including inappropriate material such as marriage proposals in articles[6]. This sort of thing is easily checked for in a program like `ckms` although on occasion the inclusion of such words as expletives is justifiable[7].

Automated checks on mathematics

`ckms` performs a bracket consistency check: every opening bracket must be matched by a closing bracket of the same type. There are some exceptions to this such as the notation for a semi-open interval, $(a, b]$, which `ckms` is programmed to ignore.

Manual checks

There are many checks that at present cannot be automated. These are mainly to do with graphical material. To save typing, stock requests to, for example, increase the font size in graph axes, can be made at the push of a button.

AUTOMATED EDITING: fixTEX

Operation

fixTEX is run on a `.tex` file from the command line and has no graphical output. It has a large number of switches some of which can be changed from their default states via command line options. The remainder of the switches are listed in a configuration file which is searched for and the states of the switches read when the program is run and created if it is not found. An example of a switch is whether or not to add a non-breaking space (a ~ in LATEX) between any number followed by a recognized unit.

Types of correction

There are three types of correction done by fixTEX: (i) cosmetic changes do not affect the output (`.ps` or `.pdf`) file but make the code easier to read (e.g., by removing unnecessary spaces and redundant commands and expanding abbreviations of LATEX commands); (ii) automated corrections (e.g., simple errors in punctuation, spelling, and grammar, addition of non-breaking space between numbers and units, putting multi-line equations into a standard form); and (iii) user-specified changes (e.g., regular expression substitution, replacement of one symbol by another in mathematical expressions only, aligning entries in a table).

Acknowledgements: I am grateful to my fellow *ScienceAsia* associate editors for helping to road test the software over the years.

REFERENCES

1. Lamport L (1994) *LATEX: A Document Preparation System*, 2nd edn, Addison-Wesley, Reading, MA.
2. Brischoux F, Legagneux P (2009) Don't format manuscripts. *Scientist* **23(7)**, 24.
3. Knauff M, Nejasmic J (2014) An efficiency comparison of document preparation systems used in academic research and development. *PLoS ONE* **9**, e115069.
4. Knuth DE (1984) *The TEXbook*, Addison-Wesley, Reading, MA.
5. Allen MA (2010) On the current obsession with publication statistics. *Sci Asia* **36**, 1–5.
6. Nature Publishing Group (2018) From proposals to gripes: the messages that scientists sneak into their papers. *Nature* **554**, 276.
7. Vaillancourt T, Sharma A (2011) Intolerance of sexy peers: Intrasexual competition among women. *Aggress Behav* **37**, 569–77.