# *Building a PDF Document using Text for Tex Markup (txt4tex) An Overview & Guide*

**by John Redmond**

Styled by LimpidSoft

# Contents

Thanks to Python and thanks, in particular, to its highly efficient *regular expression* module. This is at the heart of *txt4tex*.

It has enabled a generic approach to text processing, so that the system can readily be adapted to other target markups (an obvious target will be one or other XML-based document). For this reason txt4tex is highly modular to simplify adaptation.

# Introduction

The idea of creating a PDF document is, for most people, intimidating because the precursor .tex document is complex and foreign-looking with masses of backslashes and curly braces[1]. The idea of the *txt4tex* system is to simplify the markup, to avoid bizarre symbols *and* to enforce consistency throughout and between documents.

To create a PDF document requires two steps:

- Conversion of the text file to a LaTeX (.tex) file (using the *txt4tex* software);
- Production of a LaTeX file to a PDF file (using the LaTeX file and free software).

## What you Need

The absolute requirement is a system with a LaTeX (or equivalent) system; this is required to transform intermediate .tex documents to the final .pdf documents. Coupled with this is the requirement of a *Python* installation and a command-line interface. The examples used below use standard Unix/Linux commands, but there are equivalents on other platforms. You will need to be able to execute the following commands, or their equivalents:

*To print a file to the screen:*

```
cat mytext.txt
```

*To copy a file to another file:*

```
cat mytext.txt > anothertext.txt
```

[1] These characters were presumably chosen for the markup *because* they are rarely encountered in text.

*To process a source file* (say, with a python script) and *save the result to another file:*

```
cat mytext.txt | python processor.py > anewtext.xyz
```

*To be more specific:*

```
cat mytext.txt | python 2tex.py sampleformat.txt > anewfile.tex
```

*Or, using Python3 (recommended):*

```
cat mytext.txt | python3 2tex3.py sampleformat.txt > anewfile.tex
```

These last examples are *exactly* the command lines for building a LaTeX document (Python executes the *2tex.py* or *2tex3.py* script with a parameter that specifies the format and styling of the output .tex document). The output is directed to *anewfile.txt*.

The final pdf document the following command is executed *twice*:

```
pdflatex anewfile OR
pdflatex anewfile.tex
```

Of course, you will need the software:

- Python (Python3 recommended). This will be already installed in Linux systems. Otherwise, download (it is free).
- A LaTeX (or equivalent) installation. This, too, is free. There are different (and often confusing) options. In Linux, use your package manager. Look first for *texlive* (and be prepared to be bewildered by the options). In my experience, the plain *TexLive* lacks some of the required features, but there are lots of add-ons, so that you can add them in stages. At the top end is *texlive-full*. Very complete, very big and slow to download. This is the package that I use.

Before we move on to a discussion of what and why, I must emphasise that it is *not necessary* for a user to understand the details in the *sampleformat.txt* file. It is steeped in LaTeX detail, and for most users, all that will be necessary will be to:

- Set the page dimensions;
- Set the font type and size;

Otherwise, you will get a document with A5-sized pages and narrow margins–a typical format used by LimpidSoft.

## PREPARING TEXT SOURCE

LimpidSoft is indebted to the site at *gutenberg.com* for the books which it has processed. That said, the texts are a real mixture, so that it is wise not to assume too much about their detailed structure. Issues that emerge include:

**CHARACTER ENCODING.** The site seems, wisely, to be moving to publishing text in UTF-8 format (this is a specification for representing characters). In the past, US-ASCII was the dominant format. Assume that you have acquired a text, not necessarily from Gutenberg. You should post-haste adapt it to UTF-8 encoding before you proceed with editing. How? By using the right editor in the right way! (I do all my text editing with *gedit*, but there are equivalents on other platforms. I recommend that you *not* use a word processor.)

First of all, load the file into the editor and *immediately* save it to another file name, but *only* after specifying *UTF-8* encoding. Then you can proceed to edit the new file to your requirements. If you tinker before saving, you may have an invisible surprise waiting for you:

**ENCODING BYTES.** These are the first two (invisible) bytes at the start of the text file. They specify the the encoding status of the document. If these bytes persist (because you did *not* save in *UTF-8* immediately) and cause any difficulty, simply ignore them by leaving the *first line blank* in the edited file. (It will be ignored by Python.)

**LINE TERMINATION.** Some of us are old enough to remember the CP/M 8-bit system that formed the basis of Windows 1. CP/M terminated text lines with the old carriage return/line feed combination (0D/0A, 13/10 or \r\n). This is archaic and unnecessary and, for current work, should be replaced with just the line-feed (newline) character (\n)–provided that your editor can cope with this change. Now, LaTeX can cope with the carriage return characters if they are used consistently but, for a cleaner way forward, remove all '\r' characters.

To adapt imported text for txt4tex processing, will need to follow the sequence:

- *Remove extraneous sections* (such as the head and tail sections of a Gutenberg document);
- *Pack each group of lines into a paragraph*, followed by *at least* one blank line (see below);
- *Check paragraph starts.* If the first character is a '.' (dot) or underscore character, add a space character before it;

- *Correct the quote characters*. Replace opening quotes with backticks (and opening double quotes with two backticks) and closing quotes with one or two ticks. LaTeX will process them appropriately.

Almost there, but there remains a structural gotcha: the confusion of lines and paragraphs. To illustrate, the following text was arbitrarily lifted from Conrad's "A Secret Agent" from *www.gutenberg.com*.

> And Mr Vladimir developed his idea from on high, with scorn and
> condescension, displaying at the same time an amount of ignorance as to
> the real aims, thoughts, and methods of the revolutionary world which
> filled the silent Mr Verloc with inward consternation. He confounded
> causes with effects more than was excusable; the most distinguished
> propagandists with impulsive bomb throwers; assumed organisation where in
> the nature of things it could not exist; spoke of the social
> revolutionary party one moment as of a perfectly disciplined army, where
> the word of chiefs was supreme, and at another as if it had been the
> loosest association of desperate brigands that ever camped in a mountain
> gorge. Once Mr Verloc had opened his mouth for a protest, but the
> raising of a shapely, large white hand arrested him. Very soon he became
> too appalled to even try to protest. He listened in a stillness of dread
> which resembled the immobility of profound attention.
>
> "A series of outrages," Mr Vladimir continued calmly, "executed here in
> this country; not only *planned* here—that would not do—they would not
> mind. Your friends could set half the Continent on fire without
> influencing the public opinion here in favour of a universal repressive
> legislation. They will not look outside their backyard here."
>
> Mr Verloc cleared his throat, but his heart failed him, and he said
> nothing.
>
> "These outrages need not be especially sanguinary," Mr Vladimir went on,
> as if delivering a scientific lecture, "but they must be sufficiently
> instance. What is the fetish of the hour that all the bourgeoisie
> recognise—eh, Mr Verloc?"

What is happening? This text, as it stands, is a sequence of lines, each terminated with a newline character. The difficulty is that LaTeX treats each input line as a *paragraph*. As we will see later, in our discussion of fonts, there is a difference between the spacing of lines and the greater spacing of paragraphs. In the present example, we have a series of lines, some of them blank, which need to be grouped into a series of multiline paragraphs. This can be done in your editor by doing two global replacements:

```
replace '\n' with a space and '\n\n' with '\n'
```

This turns out to be a little trickier than expected, so it might be best to use an txt4tex utility (*join.py*):

```
cat starttext.txt | python BIN/join.py > starttext1.txt
```

To summarise the handling of whitespace by LaTeX:

**1.** Multiple (consecutive) spaces are treated as a single space: frustration for beginners[2].

**2.** One or more newline characters are treated as a single newline–and is typically used to separate text blocks (most commonly, paragraphs).

**3.** Spaces between the lines of a paragraph are less than spaces between paragraphs. In the latex preamble of the present document is the specification;

```
\newcommand{\palatinofont}{\fontsize{11}{11.5}\usefont{T1}{ppl}{m}{n}}
```

It would be distracting to give a detailed account of this very off-putting Latex code; suffice to say that it configures Palatino for this document. (As an aside, there is a group of fonts (this one is code-named *ppl*) which are expected to be available on *any* computer platform.) Here, we have specified that the default size of print for this document will be 11 points[3], with a line spacing of 11.5 points.

In the preamble to the present document there is a specification for that *additional* space (parskip) at the ends of paragraphs:

```
4pt plus2pt minus2pt
```

This is a *rubber length* which, depending on where this skip occurs on a page, can be anything from 2pt to 6pt. This means that the paragraph space is always somewhat more than the line space within the paragraph.[4]

[2] There *are* ways around this: see the section on quotes and verse.

[3] There are approximately 72 points to the inch.

[4] This is the sort of detail that intimidates potential users into using using a word processor instead of a professional text tool! Be aware of it, but lose no sleep over it.

A final point on the management of whitespace is the treatment of line endings. Just occasionally, the stream of words will fit neatly into the width of a line, but usually will not. There is then the option of spreading the words across the page, by automatic hyphenation and adding to the whitespace between the words, or of leaving space at the right margin and moving to the next line. The default behaviour in Latex is to spread across the space (full justification), which is almost always the better choice. If, however, this is not what you want, you should specify ragged right. In the tex4tex environment, this would correspond to a single statement at the start of a line:

```
_raggedright__
```

# Document Structure

At this point, we have a text document consisting of a series of paragraphs, optionally separated by blank lines for readability. We now turn to how a non-trivial document is structured, such as into chapters, sections, subsections, etc, using *dot markers* at the *starts* of particular paragraphs. The idea is simple:

- Title: one dot, as in '.My Life Story'
- Chapters: two dots, as in '..Life at School'
- Sections: three dots, as in '...My Best Friends'
- Subsections, etc: four dots (but these should be used sparingly)

And there an elaboration of the chapters and sections, with an optional subheading of the form:

```
..CHAPTER I=THE PICKWICKIANS
```

With this option, the only *true* part of the heading is 'CHAPTER 1' (and this is what appears in the table of contents). 'THE PICKWICKIANS' is dispayed as a centered *subtitle* after 'CHAPTER 1'. The default handling of chapter starts is, by default, always on an odd-numbered (right-side) page. If the *otherpage* package is added to the format document, the left-side (other) page *may* start a chapter on an even-numbered page (this is appropriate for an ebook).

Why these options? The *CHAPTER* part is used to construct a table of contents (toc). Just how the table is constructed and displayed depends on the *length* of the longest entries: if they are all at the 'CHAPTER 3' level of simplicity the toc will likely be displayed as a table, otherwise as a list. And, if there are sections and subsection, there will be *levels* of entries in the toc. The choice is yours.

This dot notation potentially introduces a gotcha: such structure dots *must* be at the *starts* of lines. If there is a space at the start, the dots will be interpreted as an *ellipsis*. And this might occasionally be just what you want!

Incidentally, you will find that the dot markers will have a profound influence on the shape and presentation of your document. In fact, in some of the simpler documents, you need not add *any* further markup to the document! Unless further markup is added, the final document will consist of zero or more logical divisions (chapters, sections, etc), but there is scope for further structure: featured blocks, lists and tables[5].

Before we leave discussion of dots, there is an optional elaboration of the title (single dot) heading: it can take up to four fields, separated by bar ( | ) characters:

- Title, optionally with double backslashes to force line breaks;
- Author;
- Background or acknowledgement field (can be empty);
- Colour of front cover. (Colours are specified in the format file. Optional: can be missing.)

[5] In fact, the text can be converted to a PDF document without *any* markup (no dots, no anything!). But do not expect the document to be particularly elegant!.

# Text blocks

Blocks are defined by *underline* characters as:

```
_quote
 (quote content)
__quote
```

This is how a quote block is created; quite simple, but with a built-in caveat: a block starts with a line with a _single underscore at the *start of a line* and ends with a*double underscore at the start of a line*. Now, there is a convention in marked-up text to emphasise a work or phrase by surrounding it by underscores:

```
 ...this is an _emphasised_ word...
```

This convention has be incorporated into txt4tex, with another variant:

```
 ...this is a _+bolded+_ word...
```

And, of course, if you have a line that starts with such underscore markup, add a space before–otherwise it will be parsed as a block marker.

There are some important options for blocks:

1. Preformatted text–which should be used only when absolutely necessary!

2. Indented text with no special text treatment;

3. Quoted text with preservation of whitespace;

4. Versified text with preservation of whitespace and indented line continuation;

## PREFORMATTED TEXT

Using a *pre* tag (for literal text) is almost *never* a good option: there is no automatic line wrapping and, to maintain some semblance of layout, a non-proportional font is used. And the problems become worse with added indentation (using the *pre2* and *pre3* tags).

```
 _pre1
‘‘‘Cyclop,’ he said, ‘take a bowl of wine from the hand of your guest’’’
    ‘‘Out rushed with mighty noise all the winds‘‘
‘‘And straight they were transformed into swine’’
      ‘‘‘Who or what manner of man art thou?’’’

‘‘And the dead came to his banquet’’
‘‘He would have broken his bonds to rush after them’’
‘‘Nine days was he floating about with all the motions of the sea’’
‘‘Took a last leave of her and of her nymphs’’
‘‘And Nausicaa joined them in a game with the ball’’
 __pre1

 _pre2
‘‘‘Cyclop,’ he said, ‘take a bowl of wine from the hand of your guest’’’
    ‘‘Out rushed with mighty noise all the winds‘‘
‘‘And straight they were transformed into swine’’
      ‘‘‘Who or what manner of man art thou?’’’

‘‘And the dead came to his banquet’’
‘‘He would have broken his bonds to rush after them’’
‘‘Nine days was he floating about with all the motions of the sea’’
‘‘Took a last leave of her and of her nymphs’’
‘‘And Nausicaa joined them in a game with the ball’’
 __pre2

 _pre3
‘‘‘Cyclop,’ he said, ‘take a bowl of wine from the hand of your guest’’’
    ‘‘Out rushed with mighty noise all the winds‘‘
‘‘And straight they were transformed into swine’’
      ‘‘‘Who or what manner of man art thou?’’’

‘‘And the dead came to his banquet’’
‘‘He would have broken his bonds to rush after them’’
‘‘Nine days was he floating about with all the motions of the sea’’
‘‘Took a last leave of her and of her nymphs’’
‘‘And Nausicaa joined them in a game with the ball’’
 __pre3
```

with the result:

```
‘‘‘Cyclop,’ he said, ’take a bowl of wine from the hand of your guest’’’
```

```
    ‘‘Out rushed with mighty noise all the winds’’
‘‘And straight they were transformed into swine’’
      ‘‘‘Who or what manner of man art thou?’’’

‘‘And the dead came to his banquet’’
‘‘He would have broken his bonds to rush after them’’
‘‘Nine days was he floating about with all the motions of the sea’’
‘‘Took a last leave of her and of her nymphs’’
‘‘And Nausicaa joined them in a game with the ball’’
```

```
          ‘‘‘Cyclop,’ he said, ’take a bowl of wine from the hand of your guest’’’
              ‘‘Out rushed with mighty noise all the winds’’
          ‘‘And straight they were transformed into swine’’
                ‘‘‘Who or what manner of man art thou?’’’

          ‘‘And the dead came to his banquet’’
          ‘‘He would have broken his bonds to rush after them’’
          ‘‘Nine days was he floating about with all the motions of the sea’’
          ‘‘Took a last leave of her and of her nymphs’’
          ‘‘And Nausicaa joined them in a game with the ball’’
```

```
                    ‘‘‘Cyclop,’ he said, ’take a bowl of wine from the hand of your gu
                        ‘‘Out rushed with mighty noise all the winds’’
                    ‘‘And straight they were transformed into swine’’
                          ‘‘‘Who or what manner of man art thou?’’’

                    ‘‘And the dead came to his banquet’’
                    ‘‘He would have broken his bonds to rush after them’’
                    ‘‘Nine days was he floating about with all the motions of the sea’
                    ‘‘Took a last leave of her and of her nymphs’’
                    ‘‘And Nausicaa joined them in a game with the ball’’
```

## Text Blocks

Before these are examined, it should be understood that LaTeX, by default, ignores multiple space characters. There are several options available, each with indentation and font style variants:

**Indents:** indent1, indent2, indent3 with the variants with sloped font: Indent1, Indent2, Indent3. There is no special treatment for whitespace.

**Quotes:** quote1, quote2, quote3, with Quote1, Quote2, Quote3. Whitespace is preserved.

**Verse:** verse1, verse2, verse3, with Verse1, Verse2, Verse3. Whitespace is preserved.

The differences cast light on the treatment of multiple whitespace by LaTeX. For *normal* text, multiple spaces are treated as a single space, which brings a problem for starting users. How txt4tex handles it is best illustrated by setting the above verse in the different environments:
As an Indented Text (indent1 block):

> “‘Cyclop,’ he said, ’take a bowl of wine from the hand of your guest”’ “Out rushed with mighty noise all the winds” “And straight they were transformed into swine” “‘Who or what manner of man art thou?”’
>
> “And the dead came to his banquet” “He would have broken his bonds to rush after them” “Nine days was he floating about with all the motions of the sea” “Took a last leave of her and of her nymphs” “And Nausicaa joined them in a game with the ball”

As a Quoted Text (Quote2 block):

> *“‘Cyclop,’ he said, ’take a bowl of wine from the hand of your guest”’*  
> *“Out rushed with mighty noise all the winds”*  
> *“And straight they were transformed into swine”*  
> *“‘Who or what manner of man art thou?”’*
>
> *“And the dead came to his banquet”*  
> *“He would have broken his bonds to rush after them”*  
> *“Nine days was he floating about with all the motions of the sea”*  
> *“Took a last leave of her and of her nymphs”*  
> *“And Nausicaa joined them in a game with the ball”*

As an Indented Verse (verse3 block):

> “‘Cyclop,’ he said, ’take a bowl of wine from the hand of your guest”’  
> “Out rushed with mighty noise all the winds”  
> “And straight they were transformed into swine”  
> “‘Who or what manner of man art thou?”’
>
> “And the dead came to his banquet”  
> “He would have broken his bonds to rush after them”  
> “Nine days was he floating about with all the motions of the sea”  
> “Took a last leave of her and of her nymphs”  
> “And Nausicaa joined them in a game with the ball”

As an Indented and Featured Verse (Verse3 block):

*"'Cyclop,' he said, 'take a bowl of wine from the hand of your guest"'*  
*"Out rushed with mighty noise all the winds"*  
*"And straight they were transformed into swine"*  
*"'Who or what manner of man art thou?"'*

*"And the dead came to his banquet"*  
*"He would have broken his bonds to rush after them"*  
*"Nine days was he floating about with all the motions of the sea"*  
*"Took a last leave of her and of her nymphs"*  
*"And Nausicaa joined them in a game with the ball"*

As a Centered and Featured Block:

*"'Cyclop,' he said, 'take a bowl of wine from the hand of your guest"'*  
*"Out rushed with mighty noise all the winds"*  
*"And straight they were transformed into swine"*  
*"'Who or what manner of man art thou?"'*

*"And the dead came to his banquet"*  
*"He would have broken his bonds to rush after them"*  
*"Nine days was he floating about with all the motions of the sea"*  
*"Took a last leave of her and of her nymphs"*  
*"And Nausicaa joined them in a game with the ball"*

# Lists, Tables and Imported Content

## Lists

The simplest option is to use preformatted text (*pre1* block: again disparaged):

```
Title: Use one dot, as in ’.My Life Story’
Chapters: Use two dots, as in ’..Life at School’
Sections: Use three dots, as in ’...My Best Friends’
Subsections, etc: Use four dots (but these should be used sparingly)
```

As an unordered list (*ul* block):

- Title: Use one dot, as in ′.My Life Story′
- Chapters: Use two dots, as in ′..Life at School′
- Sections: Use three dots, as in ′...My Best Friends′
- Subsections, etc: Use four dots (but these should be used sparingly)

As an ordered list (*ol* block):

1. Title: Use one dot, as in ′.My Life Story′
2. Chapters: Use two dots, as in ′..Life at School′
3. Sections: Use three dots, as in ′...My Best Friends′
4. Subsections, etc: Use four dots (but these should be used sparingly)

As a description list (*dl* block):

**Title:** Use one dot, as in ′.My Life Story′

**Chapters:** Use two dots, as in '..Life at School'

**Sections:** Use three dots, as in '...My Best Friends'

**Subsections, etc.:** Use four dots (but these should be used sparingly)

This last block uses the following markup:

```
_dl
Title:|Use one dot, as in '.My Life Story'
Chapters:|Use two dots, as in '..Life at School'
Sections:|Use three dots, as in '...My Best Friends'
Subsections, etc.:|Use four dots (but these should be used sparingly)
__dl
```

(Note the use of a bar (|) character here to separate the two fields. The same delimiter is used elsewhere, such as to separate columns in a table.)

## TABLES

The following content was extracted from Hugo's *Les Miserables*. It illustrates the layout of non-trivial tabular material:

```
_table|0.9|LRC|NOTE ON THE REGULATION OF MY HOUSEHOLD EXPENSES
For the little seminary|1,500|livres
Society of the  mission|100|''
For the Lazarists of Montdidier|100|   ''
Seminary for foreign missions in Paris|200|   ''
Congregation of the Holy Spirit|150   |''
Religious establishments of the Holy Land|100  |''
Charitable maternity societies|300   |''
Extra, for that of Arles|50   |''
Work for the amelioration of prisons|400  | ''
Work for the relief and delivery of prisoners|500|   ''
To liberate fathers of families incarcerated for debt|1,000|''
Addition to the salary of the poor teachers of the diocese|2,000|   ''
Public granary of the Hautes-Alpes|100|   ''
Congregation of the ladies of D----, of Manosque, and of Sisteron,\\for the gratuitous
For the poor|6,000|   ''
My personal expenses|1,000|   ''

Total|15,000|   ''

__table
```

The result appears on the following page. Note that, here, we have a more complex start tag to the table, with fields delimited by the bar ('|') character as in a definition list. In order, the line specifies that the table width is to be 0.9 times the document text width, that the alignments of the three columns are to be *left, right, center*, and that the table title is to be 'NOTE ON...'.

# NOTE ON THE REGULATION OF MY HOUSEHOLD EXPENSES

| | | |
|---|---:|:---:|
| For the little seminary | 1,500 | livres |
| Society of the mission | 100 | ″ |
| For the Lazarists of Montdidier | 100 | ″ |
| Seminary for foreign missions in Paris | 200 | ″ |
| Congregation of the Holy Spirit | 150 | ″ |
| Religious establishments of the Holy Land | 100 | ″ |
| Charitable maternity societies | 300 | ″ |
| Extra, for that of Arles | 50 | ″ |
| Work for the amelioration of prisons | 400 | ″ |
| Work for the relief and delivery of prisoners | 500 | ″ |
| To liberate fathers of families incarcerated for debt | 1,000 | ″ |
| Addition to the salary of the poor teachers of the diocese | 2,000 | ″ |
| Public granary of the Hautes-Alpes | 100 | ″ |
| Congregation of the ladies of D——, of Manosque, and of Sisteron, for the gratuitous instruction of poor girls | 1,500 | ″ |
| For the poor | 6,000 | ″ |
| My personal expenses | 1,000 | ″ |
| Total | 15,000 | ″ |

As with the addition of images (discussed below), tables need to be placed carefully in light of page breaks. If this limitation is not attractive, or if you have an occasional long table (one that *cannot* fit onto a single page) you can specify a *longtable*. It is specified in almost the same way, except that you cannot control the width. As with the *table* setting, long fields in a *longtable* can be split over multiple lines by using a double backslash. Whitespace is ignored, so that it is not possible to pad the content (so rely on the alignments given in the command). Furthermore, it is essentially a fairly primitive device for displaying literal text; if a line is too long, it disappears off the right side of the page.

If you find yourself with a table that cannot be accommodated on a page, you may be well advised to divide it into two or more logical subtables. This would avoid the horrid *longtable* and probably help the reader to understand the table content!

```
_longtable|LRC|NOTE ON THE REGULATION OF MY HOUSEHOLD EXPENSES
(content as before)
```

# NOTE ON THE REGULATION OF MY HOUSEHOLD EXPENSES

| | | |
|---|---|---|
| For the little seminary | 1,500 | livres |
| Society of the mission | 100 | ″ |
| For the Lazarists of Montdidier | 100 | ″ |
| Seminary for foreign missions in Paris | 200 | ″ |
| Congregation of the Holy Spirit | 150 | ″ |
| Religious establishments of the Holy Land | 100 | ″ |
| Charitable maternity societies | 300 | ″ |
| Extra, for that of Arles | 50 | ″ |
| Work for the amelioration of prisons | 400 | ″ |
| Work for the relief and delivery of prisoners | 500 | ″ |
| To liberate fathers of families incarcerated for debt | 1,000 | ″ |
| Addition to the salary of the poor teachers of the diocese | 2,000 | ″ |
| Public granary of the Hautes-Alpes | 100 | ″ |
| Congregation of the ladies of D—-, of Manosque, and of Sisteron, for the gratuitous instruction of poor girls | 1,500 | ″ |
| For the poor | 6,000 | ″ |
| My personal expenses | 1,000 | ″ |
| Total | 15,000 | ″ |

## Imported Content

The *txt4tex* system has distinct limitations. It was devised to handle what could be considered to be *normal* text content. It can, however, be a useful tool for providing context for *technical* content, e.g. mathematical expressions, by using *import* blocks to encapsulate *valid* LaTeX code:

```
_import

$\int_{a}^b = (b^3 - a^3)/3$

$\sum_{j=1}^n j^2 = (n+1)(2n+1)$

\[\int_{a}^b = \frac{(b^3 - a^3)}{3}\]

\[\sum_{j=1}^n j^2 = (n+1)(2n+1)\]

__import
```

The content inside the *pre* block is added to the current document *without processing* to give:

$\int_{a}^b = (b^3 - a^3)/3$
$\sum_{j=1}^n j^2 = (n+1)(2n+1)$

$$\int_{a}^b = \frac{(b^3 - a^3)}{3}$$

$$\sum_{j=1}^{n} j^2 = (n+1)(2n+1)$$

If you have read to this point, you should consider getting a formal text on LaTeX. There are several books, but probably the most popular is *Guide to LaTeX (Fourth Edition)* by Helmut Kopka and Patrick Daly.

# Consistency of Styling

There are often recurrent style patterns in a document, so that it makes sense to standardise them into a set of markup patterns to give styling consistency throughout the document–and between related documents. Apart from being consistent, these patterns can be adapted–simply by changing the definitions in the particular preamble. And, of course, related documents can be reprocessed to new styles.

I have found, for example, the '!::' and '!:' tags to be very useful in styling plays; the start of each line/paragraph is the speaker, as in:

```
!:SVIETLOVIDOFF. My audience has gone home. They are all asleep,
 and have forgotten their old clown. No, nobody needs me, nobody
 loves me; I have no wife, no children.
```

which yields:

**SVIETLOVIDOFF.** My audience has gone home. They are all asleep, and have forgotten their old clown. No, nobody needs me, nobody loves me; I have no wife, no children.

Internally, the emphasised 'SVIETLOVIDOFF' is styled as specified in the document preamble. This means that all lines starting with '!:' or '!::' will always be styled in the same way (in this case, the style is 'Emph', a specialisation of the standard 'emph' style). If the style does not suit, all that is required is an edit of the format file to redefine 'Emph'. As a result, the whole document is restyled! And, of course, all documents processed with the same format file will be restyled in the same way.

More broadly, the line stylers are:

```
+--Addition to the salary of the poor teachers of the diocese
!::Addition to the salary: of the poor teachers of the diocese
!:Addition to the salary of the poor teachers of the diocese
!+Addition to the salary of the poor teachers of the diocese
!!Addition to the salary of the poor teachers of the diocese
<<Addition to the salary of the poor teachers of the diocese
>>Flush right addition to the salary of the poor teachers of the diocese
```

with the result:

ADDITION to the salary of the poor teachers of the diocese and with some extra text added to force text overflow onto the next line. This is because, to be convincing, a full drop cap occupies the height of a full *three* text lines!

**ADDITION** to the salary of the poor teachers of the diocese

**ADDITION TO THE SALARY!** of the poor teachers of the diocese

**ADDITION** to the salary of the poor teachers of the diocese

**Addition to the salary of the poor teachers of the diocese**

*Addition to the salary of the poor teachers of the diocese*

Addition to the salary of the poor teachers of the diocese

Flush right addition to the salary of the poor teachers of the diocese

Two final, subtle points:

Featured lines can be included in an indent or quote block (as here), with two exceptions:

1. ’>>’ works only at the document top level, forcing the text across to the right margin;
2. And the ’<<’ markup is intended to cancel any automatic indentation, such as at the starts of paragraphs. (Recall that the first paragraph in a chapter or section is normally *not* indented, whereas any later paragraphs *are* indented–unless they start with ’<<’[6].)

### TEXT COLUMNS

All or part of a document can be displayed in columns as:

```
_twocolumns__ and _threecolums__
   and cancelled by _onecolumn__
 or cancelled automatically at the end
     of the document (not recommended!!)
```

The following text block is displayed in three columns–just to show that we can!

In this document, we have glanced at some of the more complex aspects of markup; there is much, much more and the best advice I can offer to anyone who has read this far is to look through the books at LimpidSoft. There you will find the marked-up text of all the PDF books on the site, ranging from Wodehouse froth to complex tomes from Montaigne and Pepys.

[6] If indentation is *not* wanted throughout the document, alter the *parindent* setting in the format document to 0ex.

# Declarations, Leaves and Images

## Declarations

In txt4tex parlance, these are *empty statements*, aligned at the starts of lines, commencing with a single underscore and ending with two underscores:

```
_nextpage__: fill the remainder of the current page with
   whitespace and insert a page break;

_later__: insert a tiny graphic to indicate a break in
   text continuity;
_donumbering__: commence numbering of lines of verse and
   marking every tenth line;
_nonumbering__; stop numbering

_onecolumn__: revert to full page spread;
_twocolumns__: start laying out of text into two columns;
_threecolumns__: start laying out of text into three columns;
_endcolumns__: stop laying out of text into two columns;
_theendnotes__: list the accumulated endnotes at this point
  in the document.
```

There is also a specialist group of declarations to influence the layout of the *table of contents*. Using the count of headings and subheadings, and their length, the software attempts to optimise this, but it is possible to intervene with declarations:

```
_notoc__: no table of contents,
          even if there are appropriate headings;
_toccolumns__: insist on a table of contents,
          and place it _here_ in the document;
_frametoc__: place the table of contents
          in the center (centre) of the page.
```

## Leaves

These are generally simple one-liners, but have diverse rôles:

```
_lang=german, greek, italian, etc
_vspace=3 (or any other number)
_squeeze=3 (or any other number)
herepic, floatpic, centerpic, leftpic, or rightpic
```

The *lang* leaf influences some of the fine detail of the final document and some of the more obvious markup. It is good general policy to include greek and english in this group. As an example:

```
_lang=greek, english, german
```

As a result. the final book automatically has an *Inhaltsverzeichnis* instead of a *Contents* section. The inclusion of *greek* and *english* averts any difficulties with references and quotations–and, of course, we could have added other languages; the last language in the list becomes the language of the final book.

The *vspace* leaf is used to place images or blocks precisely on a page, as on the front page of a *LimpidSoft* book. The vspace is allocated on a relative basis:

```
_vspace=2
_herepic=LOGO.png|0.6
_vspace=3
```

And there are other variants on images and their management:

### Images

Figure 1: This is a Float Pic.

The management of images and their placement is very much trial and error. If an image is too big or in the wrong position, LaTeX will throw a page break and give you a very disappointing result. Part of the problem is that it is impossible, in any elegant way, to anticipate *where* the page breaks will fall. And the image may be too big or too small. Tinkering with the image size can simplify things–and it is generally useful to ensure that the vertical size of the image is not greater than the size of the block it is associated with.

Broadly, the leftpic and rightpic images are placed where you intended, provided that there is not a page break looming, while the floatpic images will go where LaTeX decides to put them! In extreme cases, where you are attempting to crowd left or right images, you may find that one of the images is missing from the final document! The solution is to change the image position in relation to the adjacent text, and this may help to avoid excessive line truncation if the image protrudes past the current paragraph.

Each of these image types can process an optional scaling factor. Fix this first, inevitably by trial and error, so that the image will fit comfortably on *a* page. Your placement options will be determined by the position of the image in the desired page breaks–for example, if the image is at the start of a chapter or section, you can use a center, left or right pic. But in other cases–and really in *all* cases, a *floatpic* is the safest and easiest because LaTeX will determine, at the last point in the layout, where the best position is.

Figure 2: This is a Right Pic.

To get down to the detail of the image specifications with examples. To illustrate typical use, most of the images are placed in some irrelevant text.

```
_centerpic=tess.jpg|0.8|This is a Center Pic.
_floatpic=tess.jpg|0.8|This is a Float Pic.|ht
_leftpic=tess.jpg|0.5|A Left Pic.|0.4
_rightpic=tess.jpg|0.5|A Right Pic.|0.4
```

**CENTERPIC:** The simpest (and most risky) option with a scaling factor (to reduce the size of the image) and a caption. The image will be centered on the page in the *present* position on the page. If there is not enough space at that position, it will be moved to a new page, leaving whitespace ahead of it. Better avoided, particularly because the contents of the unfilled previous page fill be *floated* with unseemly gaps. If the use of centerpic is unavoidable, a *nextpage* command should be inserted before it.

**FLOATPIC:** Similar to the centerpic, but with the advantage of having some of the following text moved automatically ahead of the image in order to avoid the whitespace. The *ht* specification directs the image to *here* if there is space, otherwise to the *top* of the next page. A better option.

**LEFTPIC OR RIGHTPIC** These include the scaling factor, as before, and a factor to scale the horizontal space allocated to the image. This extra scaling factor is the *reduction* of the text width to the right or left of the image.

*Start of Irrelevant Text.*

It is characteristic of this buoyant people that they pursue no man beyond the grave. “Let God be his judge¡‘–Even with the hundred thousand unfound, though greatly coveted, the hue and cry went no further than that.

To the stranger or the guest the people of Coralio will relate the story of the tragic end of their former president; how he strove to escape from the country with the public funds and also with Doña Isabel Guilbert, the young American opera singer; and how, being apprehended by members of the opposing political party in Coralio, he shot himself through the head rather than give up the funds, and, in consequence, the Señorita Guilbert. They will relate further that Doña Isabel, her adventurous bark of fortune shoaled by the simultaneous loss of her distinguished admirer and the souvenir hundred thousand, dropped anchor on this stagnant coast, awaiting a rising tide. They say, in Coralio, that she found a prompt and prosperous tide in the form of Frank Goodwin, an American resident of the town, an investor who had grown wealthy by dealing in the products of the country–a banana king, a rubber prince, a sarsaparilla, indigo, and mahogany baron. The Señorita Guilbert, you will be told, married Señor Goodwin one month after the president's death, thus, in the very moment when Fortune had ceased to smile, wresting from her a gift greater than the prize withdrawn. Of the American, Don Frank Goodwin, and of his wife the natives have nothing but good to say. Don Frank has lived among them for years, and has compelled their respect. His lady is easily queen of what social life the sober coast affords.

Figure 3: A Left Pic.

The wife of the governor of the district, herself, who was of the proud Castilian family of Monteleon y Dolorosa de los Santos y Mendez, feels honoured to unfold her napkin with olive-hued, ringed hands at the table of Señora Goodwin. Were you to refer (with your northern prejudices) to the vivacious past of Mrs. Goodwin when her audacious and gleeful abandon in light opera captured the mature president's fancy, or to her share in that statesman's downfall and malfeasance, the Latin shrug of the shoulder would be your only answer and rebuttal. What prejudices there were in Coralio concerning Señora Goodwin seemed now to be in her favour, whatever they had been in the past.

It would seem that the story is ended, instead of begun; that the close of a tragedy and the climax of a romance have covered the ground of interest; but, to the more curious reader it shall be some slight instruction to trace the close threads that underlie the ingenuous web of circumstances.

To the stranger or the guest the people of Coralio will relate the story of the tragic end of their former president; how he strove to escape from the country with the public funds and also with Doña Isabel Guilbert, the young American opera singer; and how, being apprehended by members of the opposing political party in Coralio, he shot himself through the head rather than give up the funds, and, in consequence, the Señorita Guilbert. They will relate further that Doña Isabel, her adventurous bark of fortune shoaled by the simultaneous loss of her distinguished admirer and the souvenir hundred thousand, dropped anchor on this stagnant coast, awaiting a rising tide. They say, in Coralio, that she found a prompt and prosperous tide in the form of Frank Goodwin, an American resident of the town, an investor who had grown wealthy by dealing in

Figure 4: This is a Center Pic.

the products of the country–a banana king, a rubber prince, a sarsaparilla, indigo, and mahogany baron. The Señorita Guilbert, you will be told, married Señor Goodwin one month after the president's death, thus, in the very moment when Fortune had ceased to smile, wresting from her a gift greater than the prize withdrawn.

*End of Irrelevant Text.*

Note that these image commands are simply examples of leaves, but have been complicated by fields separated by '|' characters. To explain, the floatpic references an image file *tess.jpg*, to be scaled to 0.8 of its size and given a caption 'This is a Float Pic'. LaTeX is requested to place the image at the *here* position in the final document, otherwise at the *top* of the current page, otherwise whatever is possible. A floatpic option is always the best and should be tried on the first pass through LaTeX. Thereafter, expect to tinker, especially for the left and right images. Sometimes, it is sufficient to tweek the display size (parameter 2) and allocated space (parameter 4), remember that LaTeX rules... And, perhaps to add disappointment to difficulty, if the page dimensions are altered, the images may migrate! But where *are* the image files to be found? Regardless of the current (operating) directory, it will be assumed that all the included image files are in an immediate 'IMAGES' subdirectory. For example, if the current subdirectory is 'WORK', all the image files will be in 'WORK/IMAGES'. In the event that a required image file is not found during processing, a 'failed to find' line is inserted in the final document.

The best and final advice on images: insert markers in the text and ignore images until everything is set to your satisfaction. When you start to introduce images, set them with *floatpic* to give you a feel for how they relate to the text, and then (if you wish to) change from float another style *one image at a time*, tweaking the image sizes and, if necessary, the left or right allocated space. And be aware that the layout of images will be *very dependent* on the dimensions of the text and any paragraph breaks, which in turn are determined by *both* the page size *and* the margins you have chosen. Potentially bothersome, but this will the case for any document layout system!

# GETTING STARTED

Before you can do anything, it is essential that you have a LaTeX system installed on your computer. The best option is probably to install TexLive, but you will be embarrassed by choices. Linux provides a fairly painless way to install, by using the package manager. Specify *texlive* and look at the options. There is the basic *texlive* option, so this is a good place to start because other modules can be added later. (Just what you will need will become more obvious when you start working.) There may be the temptation to install *texlive-full*, which I have done, but be aware that it is a *very* big package and the download is likely to take upwards of *an hour*!

Unsurprisingly, what you *do* really need will depend on what you plan to do but, if you have the space, why not? And you will be able to cope with (almost) any requirement. But there is the the occasional package that is not free of licence restrictions. An example that has disappointed me is the *picins* package for adding side images in the flow of text.

Download the txt4tex.zip file and unpack it in your home directory. This will have the structure:

*WORK*
*BIN*
  *2tex.py, 2tex3.py and related files*
*FORMAT*
  *testformattxt*
  *A5format.txt*
*SOURCE*
  *txt4tex.txt*
*DEST (for the intermediate .tex files)*
*IMAGES*

You will find the marked-up text for this document in the *SOURCE* directory; so you should start to process it:

First, cd to the *txt4tex/WORK* directory and then execute the three command lines:

```
cat SOURCE/txt4tex.txt|python3 BIN/2tex3.py FORMAT/A5format.txt
   > TEX/txt4tex.tex (ALL IN ONE LINE)

pdflatex TEX/txt4tex
pdflatex TEX/txt4tex
```

Note that we are using python3 here, with the matching 2tex3.py. If you do not yet have python3 installed, use:

```
cat SOURCE/txt4tex.txt|python BIN/2tex.py FORMAT/A5format.txt
   > TEX/txt4tex.tex

pdflatex TEX/txt4tex
pdflatex TEX/txt4tex
```

The result is the file *txt4tex.pdf* in the *WORK* directory, together with a mess of intermediate files. For Linux users, I have included a clean-up Linux script (*cleanlatex*) to remove them from the *current* directory:

```
cleanlatex
```

In other environments, from time to time, you will need to delete all files with *.out*, *.log*, *.aux* and *.tex* extensions from your *work* directory.

Before we leave this section, we should note that there is an *optional* embellishment to the first of the above command lines:

```
cat SOURCE/txt4tex.txt|python3/BIN/2tex3.py FORMAT/A5format.txt+0.5
   > TEX/txt4tex.tex (ALL IN ONE LINE)
```

Here, we have added a scaling factor to the format specifier, specifying that *all* included graphics are to be scaled by a factor of 0.5. This has proved a very useful option in processing text to documents of different shapes and sizes[7]. But, for most purposes, this will not be needed.

> This has proved a very useful option in processing text to documents of different shapes and sizes

### Footnotes and Endnotes

We have just encountered a footnote. This was inserted at the end of a normal text sentence[8] Note how the double-at characters can be separated by empty braces to block processing them as the start or end of a note:[9].

If you use this option, you need to declare *where* you want the note (or group of notes). This is done by a declaration:

[7] Compare the pages for an iPad and for a Kindle. There *is* a difference!.

[8] A footnote or endnote need not be at the end of a sentence. It can be anywhere and delimited by '@@' at the start and end of the note. The task of how to add the footnote is left to LaTeX, but it is *always* on the same page. A less common option is to use an *endnote*. This is delimited by '@@end=' at the start.

[9] Using syntax to discuss syntax (as here) can become somewhat convoluted! This paragraph is ended with an end note[1] Endnotes can be a practical alternative to footnotes if you are working with a documents has proved a very useful option in processing text for documents with small page sizes.

```
        _theendnotes__
```

This allows you to collect the endnotes at several points at (say) the ends of chapters or at the end of the document. (The *NOTES* heading is inserted automatically by LaTeX.)

## NOTES

[1]This is the substance of the end note, which is a good device for including longer comments–or extended blocks of illustrative waffle. Used systematically, endnotes may occasionally be more suitable than footnotes. And, to help the reader, endnotes are best at the ends of logical sections–*or* at the end of the book.

### THE DOCUMENT PREAMBLE

I have left discussion of the LaTeX preamble to the last because it is the most intimidating part of using the system. Broadly, incidental users need know little more than the fact that a preamble is needed–and that the preamble is the main part of the format file.

The preamble *is* essential because it specifies all the components of LaTeX that are needed for *this* particular document. It specifies basic stuff, such as the shape of the page for the final document–and what margins are required. And what font is to be used, the size of the font, the space between the lines and between paragraphs (they *are* different). Try altering them! Towards the end of the preamble are several definitions of styles used in the document. Any of these can be altered–often with profound changes in the final document. A slightly complicated example is (note those backslashes and curly braces):

```
\newcommand\headingB[1]{\begin{center}\thispagestyle{empty}\large\bfseries\scshape
    #1\end{center}\vspace{2ex}}
```

A headingB heading will have the following characteristics:

1. It will be centered across the page;

2. This page will be empty (no headers or footers);

3. The size of the font will be *large* (not *normalsize* or *Large*);

4. The heading will be *bold* and in *small caps* shape.

5. There will be vertical space, equivalent to two blank lines, after the heading.

But the practical point is that the user of this preamble does not need to know or particularly care about this detail–except at the level of deciding to use this particular format.

Once you have a preamble document, *always* make a copy. Tinker with the copy and try building new documents with the copy. When you become more confident, try altering the 'hooks for featured content'.

## Selection of Fonts

The appearance and readability of a document can be greatly influenced by the font used. Fonts, their design and management is a very big part of the LaTeX background, but we will deal with it at a pretty superficial level. There is a group of fonts that can be assumed available in all current computer environments. These are the so-called *Cork fonts*, and they are declared in our preambles:

```
\newcommand{\palatinofont}{\fontsize{13}{14}\usefont{T1}{ppl}{m}{n}}
\newcommand{\bookmanfont}{\fontsize{10}{11.5}\usefont{T1}{pbk}{m}{n}}
\newcommand{\helveticafont}{\fontsize{10}{11.5}\usefont{T1}{phv}{m}{n}}
\newcommand{\chanceryfont}{\fontsize{9}{10.5}\usefont{T1}{pzc}{m}{il}}
\newcommand{\courierfont}{\fontsize{10}{11.5}\usefont{T1}{pcr}{m}{n}}
\newcommand{\charterfont}{\fontsize{10}{11.5}\usefont{T1}{pch}{m}{n}}
\newcommand{\timesfont}{\fontsize{12.5}{14}\usefont{T1}{ptm}{m}{n}}
\newcommand{\avantgardefont}{\fontsize{9}{10.5}\usefont{T1}{pag}{m}{n}}
```

The only part that you need understand is the size specification, eg. for Palatino, specifies a height of 13 points on a background of 14 points (this is how the space between the lines is set). The actual size used in the document is determined *indirectly*:

```
by direct control: \newcommand{\basefont}{\fontsize{10}{10.5}}
or by indirect control: \newcommand{\basefont}{\smallfont}
  (it would otherwise be \normalfont).
```

The traditional approach to font management is to use one of the built-in fonts. Palatino and Times are there, of course, but it is interesting to investigate less common options, such as cmbright, inconsolata, librebaskerville and newcentury. To include one of them, e.g. cmbright, add the following to the preamble[10]:

```
\import{cmbright}
\newcommand{\basefont}{\fontsize{10}{10.5}}
```

## Page Breaks

You will find that there is an ever-present potential problem–bad page breaks. The problem here is that LaTeX generates a page break when the current page is full. This can result in a small number of orphan lines on the new page. A partial solution might

[10] A final caveat: we are looking at two versions of the basefont command here: you can have *one* or *neither*, but never *two* of them.

be to throw a page break (with the *nextpage* command slightly before the automatic break (this has immediate effect after the current line). Alternatively, attempt to delay the page break (by a line or two) with the *squeeze* command[11]. To illustrate:

```
_nextpage__

The best and final advice on images: insert markers...
_nextpage__
And be aware that the layout of images will be...

OR

_squeeze=2
The best and final advice on images: insert markers...
...but this will the case for any document layout system!
```

## Licensing

> ***This program package is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.***

So... You can test the waters without risk. I suggest that you *not* be distracted by the detail in this file (because it is artificially complicated). Better to look at the books on the LimpidSoft site: find a few that resemble a document that you might want to process and *copy, copy, copy*!

## Coda

Is the txt4tex system perfect? Certainly *not!* It may, however, be a step in the right direction for improving document quality–perhaps especially for *ebooks*.

It is instructive to explore the limits of the practical. To take a complex document like txt4tex.txt and expect it to adapt seemlessly to, say, to Kindle-sized small pages will disappoint. Of course, this can be done–and *has* been done, but may not really be worth doing.

John Redmond
Sydney, Australia.

[11] This seems to work best if the squeeze declaration is placed further up the page. Note that it is a *request* to LaTeX, and it may not give the desired result. Perhaps you are requesting too much extra space, so that you get less–or even none.