Incipit:
A Unicode-based Text Markup Language

Seninha (aka phillbush)

The *Incipit Markup Language* (or *Incipit*, for short) is a plain text markup language that uses Unicode characters and the structure of the text itself to format documents.

In the *Incipit Markup Language*, a paragraph is a block of text delimited by blank lines. A paragraph may be preceded by a section header and succeded by a figure. Enumerations (also knonw as "*lists*") are special types of paragraphs (although some people interpret an enumeration as the continuation of the paragraph before it).

In this document, the word "*period*" refers sentences delimited by a period. And the word "*colon*" refers to a segment of text delimited by a colon or semi-colon.

1. Punctuation

Incipit uses Unicode characters (called "*punctuation*" in this document), alongside the structure of the text, to format documents. For example, the section character (`'§'`, U+00A7) is used to markup section headers. The bullet character (`'•'`, U+2022) is used to markup bulleted lists.

Inline punctuation. Punctuations are mostly used within a paragraph. Those punctuations (called inline punctuation) markup emphases, references, topics or preformated text. Punctuations cannot be nested: a portion of text is either emphasized, or it is preformated, never both. The types of inline punctuations are enumerated below.

- **Emphasis:** Text *between single quotes* is emphasized. The single quotes must be Unicode characters `''` (U+2018) and `''` (U+2019). Emphatic text is formated in italic font and the punctuation is removed in the final document.
- **Topic:** Text "*between double quotes*" is topicalized. The double quotes must be Unicode characters `''` (U+201C) and `''` (U+201D). Topical text is formated in italic font and the punctuation is preserved in the final document.
- **Reference:** Text «between double angle quotes» is reference. The angle quotes must be Unicode characters `'«'` (U+00AB) and `'»'` (U+00BB). References are not supported yet.
- **Preformated:** Text `between grave accents` or `between curly braces` is preformated. Those punctuation are regular ASCII punctuation. Preformated text is formated in monospaced font and the punctuation is removed in the final document.
- **Meta text:** Text `'between angle braces'` is meta text. It is formated in monospaced font and the punctuation is kept in the final document.

Typing punctuation. If you use Unix, you can either configure your keybindings or configure the Compose key to insert punctuation and other characters not found on a regular keyboard.

1.1. Sections

Paragraphs can be grouped in sections, which can be nested. A section is a line whose first characters are section punctuations (§, U+00A7). A section can be marked by a one or more section punctuations. The number of time that punctuation occurs represents the level of the section. For example, a first-level section begins with §; a second-level section begins with §§, and so on.

1.2. Enumerations

An enumeration, also known as list, is a hierarchical grouping of periods, called the enumeration items. Each item begins with zero or more tab characters followed the enumeration punctuation (•, U+2022), also knon as *bullet*. The number of tabs in the beginning of an item identifies the item level: zero tab for first-level items; one tab for second-level items; and so on.

Enumeration label. When formated, each enumeration item is usually preceded by a bullet. However, it can be changed by following the enumeration punctuation by a string between parentheses. This can be used for ordered lists, when the label is a number or letter.

Enumeration incipit. Each enumeration item can have a incipit colon, which will be explained on the "*§ Incipit*" section below. The incipit colon is a colon describing the topic of the item.

The following is an example of enumeration.

A. First item: This is the first item of a labeled enumeration. This item also contains an incipit colon.

B. Second item: This is the second item of a labeled enumeration. It also contains an incipit colon.

C. Third item.

- First subitem of third item.
- Second subitem of third item.
- Third subitem of third item.
- Fourth subitem of third item.

D. Fourth item.

E. Fifth item.

2. Incipit

The word "*incipit*" comes from the Latin and means "*it begins*". The incipit of a text is the first few words of the text. In the *Incipit Markup Language*, incipits are initial elements of the text used to format the text itself. The incipit of a document is its first paragraph (which contains the title and some meta information); the incipit of a paragraph is its first period (aka sentence); the incipit of a period is its first colon (the part separated by colon).

In the *Incipit* Markup Language, a text unit can have no incipit. A document without incipit is a document without title. A paragraph without incipit is a paragraph without its special first period. This implies that certain units of text are made up of two parts: an optional incipit and a body.

The incipit of a document. The first paragraph of a document is its incipit. If the document begins with a blank line or with a figure or enumeration, the document has no incipit. The first period (ie', the first sentence) of the document's incipit is the title. If this period has a incipit colon, this colon is the main title and the rest is the subtitle. For example, this document has an incipit paragraph, which has an incipit period (the full title), which has an incipit colon (the main title). The remaining periods are interpreted depending on the output format. In troff, the second period is the author, the third period is the institution, and the following periods are the abstract of the document.

The incipit of a paragraph. If the first period of a paragraph begins with a period punctuation, this paragraph contains a incipit period. This incipit period, called the title of the paragraph, is formated in bold font. In the source plain text of this document, the incipit of a paragraph is written alone in aline above the rest of the paragraph; but this is

not necessary, the incipit period can be written in the same line of the rest of the paragraph.

The incipit of a period. In an enumeration, the first colon of the first period of a enumerated item is the item's incipit colon. This incipit colon, called title of the enumeration, is formated in bold font. The first enumeration of this document, listing the types of inline punctuations, contains incipit colons.

3. Figures

Figures are text delimited between curly brackets. The opening curly bracket must be the last character in a line and the closing curly bracket must be the first character in a line. The content of a figure is usually idented with a tab, so the first tab of each code line is removed in the final document. A tag before the open curly bracket and followed by a colon indicates the type of the figure. The content between the tag and the open curly bracket is the caption

Code Listings. The simplest figure is a code listing, an example of which, copied from the second edition of "*The C Programming Language*" book, is presented below.

```
#include <stdio.h>

main()
{
     printf("hello, world\n");
}
```

Figure 1: Hello World.

PIC Diagrams. When converting to troff, diagrams can be written in the PIC language. Diagrams are marked with the `'PIC:'` keyword before the opening curly bracket. The only punctuation that are processed inside a PIC figure are emphasis and topic (topic is converted to a emphasis between ASCII double quotes).

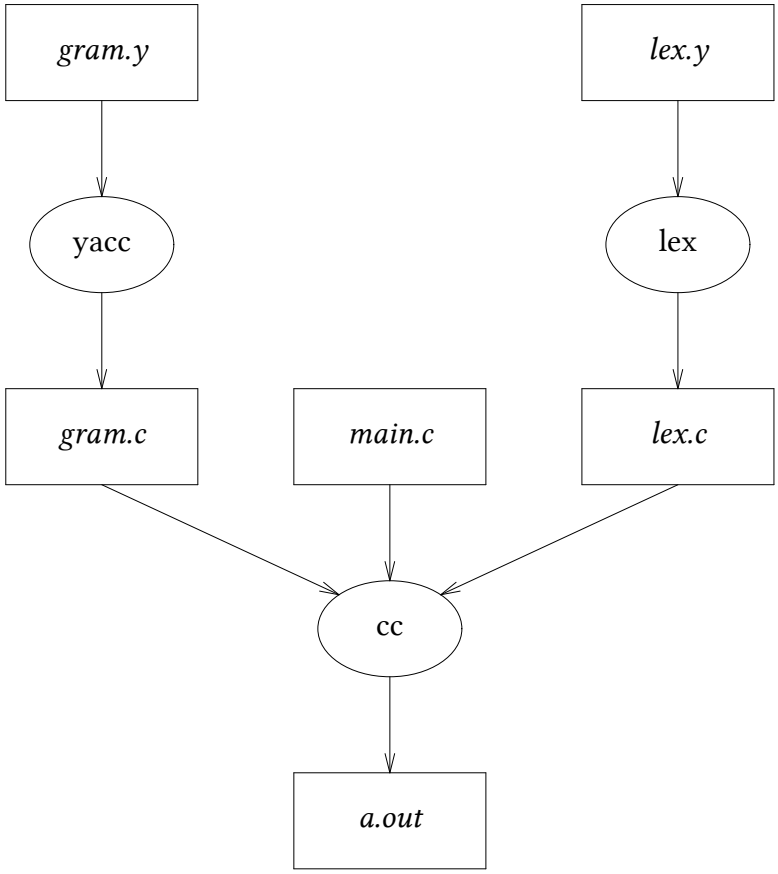

Figure 2: Compilation process.

Images. Images can be inserted on a document by preceding the opening bracket with the `'IMAGE:'` keyword, optionally followed by a caption. Note that only `.eps` images are supported when converting to troff. When converting to html, however, common formats such as `.jpg` and `.png` are supported.

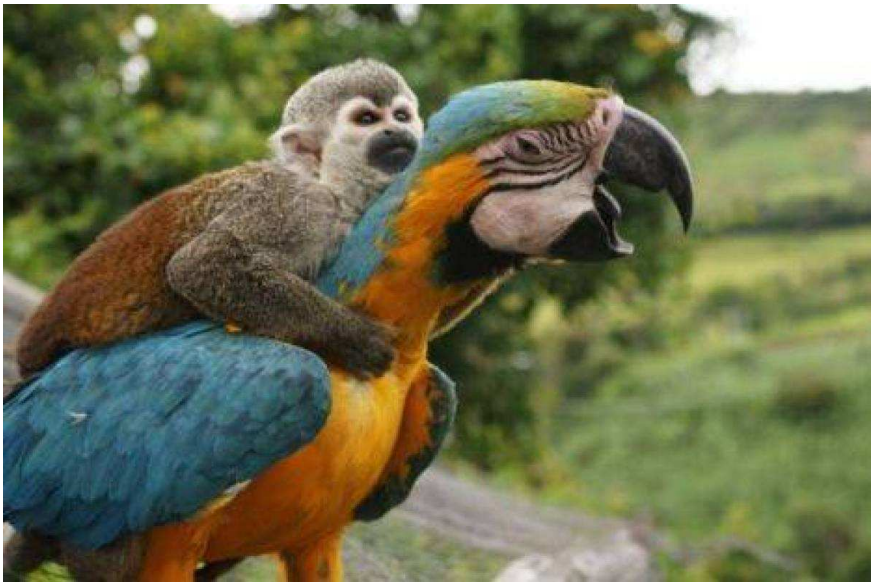

Figure 3: A monkey riding a parrot.

Tables. Tables are special figures in the sense that they are not written between curly brackets. Tables must be written using box drawing Unicode characters. Columns must be

separated by a vertical light box drawing character. The first row separator must contain double box drawing characters, and the following ones must be separated by light double box drawing characters. The following is an example of table.

COUNTRY	AREA	POPULATION	CONTINENT
Brazil	3286	134	South America
Canada	3852	25	North America
China	3705	1032	Asia
England	94	56	Europe
France	211	55	Europe
Germany	96	61	Europe
India	1267	746	Asia
Japan	144	120	Asia
Mexico	762	78	North America
USA	3615	237	North America
USSR	8649	275	Asia

Table 1: Country table from The AWK Book.

Tables (row span). If a cell in a table contains only two apostrophes (`''''`, called "*ditto*"), this cell contains the same content of the the cell above it, and both cells are merged into a single cell. We call this phenomenon a "*row span*". Row spans are only supported when converting to troff. Column span is not supported at all. The following is an example of a table with row span.

COUNTRY	AREA	POPULATION	CONTINENT
Brazil	3286	134	South America
Canada	3852	25	North America
Mexico	762	78	
USA	3615	237	
France	211	55	Europe
Germany	96	61	
England	94	56	
China	3705	1032	Asia
Japan	144	120	
India	1267	746	
USSR	8649	275	

Table 2: Country table from The AWK Book.

Tables (alternative form). *Incipit* supports an alternative format for tables, in which the first row separator contains light box drawing characters, and the following rows are not separated by any character but a new line. An example of such table is presented below.

COUNTRY	AREA	POPULATION	CONTINENT
Brazil	3286	134	South America
Canada	3852	25	North America
China	3705	1032	Asia
England	94	56	Europe
France	211	55	Europe
Germany	96	61	Europe
India	1267	746	Asia
Japan	144	120	Asia
Mexico	762	78	North America
USA	3615	237	North America
USSR	8649	275	Asia

Table 3: Country table from The AWK Book.

Quotation. Quotations are special figures which, instead of curly braces, are written between double quotes.

“
I'd just like to interject for a moment. What you're referring to as Linux, is in fact, GNU/Linux, or as I've recently taken to calling it, GNU plus Linux. Linux is not an operating system unto itself, but rather another free component of a fully functioning GNU system made useful by the GNU corelibs, shell utilities and vital system components comprising a full OS as defined by POSIX.

4. Conventions

Texts written in the Incipit markup languages use some unnusual conventions that are described in this section.

Apostrophe for abbreviations. In usual non-Incipit text, a full stop can be used both to mark the end of a period and to mark the end of an abbreviations. This ambiguity does not exist in Incipit. In Incipit, abbreviations should be marked with apostrophes.

Uppercase for emphasis. In usual non-Incipit text, the italic (or bold) font is used to render segments of text with emphasis. In Incipit, non-roman fonts are used for syntactic purposes (for example, bold text is used for the title of a paragraph, and italic is used for topics and quotations). In Incipit, text with emphasis should be written in all caps.

5. Conversion

The `i2roff(1)` and `i2html` awk scripts convert a text written in Incipit to pdf (using the `mt(7)` or `mp(7)` troff macro packages), or to HTML formats.

Converting an Incipit text to HTML is simple, just use the `i2html(1)` script to read the text file, and the converted document is printed to standard output.

```
i2html <file.txt >file.html
```

Converting an Incipit text to PDF is more complex. First, the text must be converted to `mt(7)` or `mp(7)`, which are `troff(1)` macro packages.

```
i2roff <file.txt >file.roff
```

Then, we need to convert the `.roff` file to a `.ps` (postscript) file, this is done with a series of commands in a pipeline. Those commands depend on the troff system and vary between Groff, Heirloom Doctools, etc. We need to chose one of the macro packages: `mt(7)` if we are formatting a A4 paper, or `mp(7)` if we are formatting a slide presentation in landscape letter paper. Using Heirloom Doctools, the command is the following to generate a portrait document (to generate a landscape document, used for slides, replace `portrait` with `landscape` in the command below).

```
<file.roff pic |\
tbl |\
troff -mt -mpictures |\
dpost -pportrait >file.ps
```

Then, we need to convert the `.ps` file to a `.pdf` file. This is done with the `ps2pdf(1)` command. This command accepts as argument the papersize, which can be `a4`, `letter`, `halfletter`, etc.

```
ps2pdf "-sPAPERSIZE=a4" file.ps file.pdf
```

A Makefile automatizing the conversion process is distributed with Incipit. For more information on the `mt(7)` and `mp(7)` macro packages, manuals for them are distributed as well.