GitHub Flavored Markdown Spec

Version 0.29-gfm (2019-04-06)

This formal specification is based on the <u>CommonMark Spec</u> by <u>John MacFarlane</u> and licensed under <u>(cc) BY-SA</u>

- 1 Introduction
 - 1.1 What is GitHub Flavored Markdown?
 - 1.2 What is Markdown?
 - 1.3 Why is a spec needed?
 - 1.4 About this document
- 2 Preliminaries
 - 2.1 Characters and lines
 - 2.2 Tabs
 - 2.3 Insecure characters
- 3 Blocks and inlines
 - 3.1 Precedence
 - 3.2 Container blocks and leaf blocks
- 4 Leaf blocks
 - 4.1 Thematic breaks
 - 4.2 ATX headings
 - 4.3 Setext headings
 - 4.4 Indented code blocks
 - 4.5 Fenced code blocks
 - 4.6 HTML blocks
 - 4.7 Link reference definitions
 - 4.8 Paragraphs
 - 4.9 Blank lines
 - 4.10 Tables (extension)
- 5 Container blocks
 - 5.1 Block quotes
 - 5.2 List items
 - 5.3 Task list items (extension)
 - 5.4 Lists
- 6 Inlines
 - 6.1 Backslash escapes
 - 6.2 Entity and numeric character references
 - 6.3 Code spans
 - 6.4 Emphasis and strong emphasis
 - 6.5 Strikethrough (extension)
 - 6.6 Links
 - 6.7 Images
 - 6.8 Autolinks
 - 6.9 Autolinks (extension)
 - 6.10 Raw HTML
 - 6.11 Disallowed Raw HTML (extension)
 - 6.12 Hard line breaks
 - 6.13 Soft line breaks
 - 6.14 Textual content
- Appendix: A parsing strategy
 - **Overview**
 - Phase 1: block structure
 - Phase 2: inline structure

 \blacksquare

 \mathbf{A}

1 Introduction

1.1 What is GitHub Flavored Markdown?

GitHub Flavored Markdown, often shortened as GFM, is the dialect of Markdown that is currently supported for user content on GitHub.com and GitHub Enterprise.

This formal specification, based on the CommonMark Spec, defines the syntax and semantics of this dialect.

GFM is a strict superset of CommonMark. All the features which are supported in GitHub user content and that are not specified on the original CommonMark Spec are hence known as **extensions**, and highlighted as such.

While GFM supports a wide range of inputs, it's worth noting that GitHub.com and GitHub Enterprise perform additional post-processing and sanitization after GFM is converted to HTML to ensure security and consistency of the website.

1.2 What is Markdown?

Markdown is a plain text format for writing structured documents, based on conventions for indicating formatting in email and usenet posts. It was developed by John Gruber (with help from Aaron Swartz) and released in 2004 in the form of a syntax description and a Perl script (Markdown. p1) for converting Markdown to HTML. In the next decade, dozens of implementations were developed in many languages. Some extended the original Markdown syntax with conventions for footnotes, tables, and other document elements. Some allowed Markdown documents to be rendered in formats other than HTML. Websites like Reddit, StackOverflow, and GitHub had millions of people using Markdown. And Markdown started to be used beyond the web, to author books, articles, slide shows, letters, and lecture notes.

What distinguishes Markdown from many other lightweight markup syntaxes, which are often easier to write, is its readability. As Gruber writes:

The overriding design goal for Markdown's formatting syntax is to make it as readable as possible. The idea is that a Markdown-formatted document should be publishable as-is, as plain text, without looking like it's been marked up with tags or formatting instructions. (http://daringfireball.net/projects/markdown/)

The point can be illustrated by comparing a sample of <u>AsciiDoc</u> with an equivalent sample of Markdown. Here is a sample of AsciiDoc from the AsciiDoc manual:

```
1. List item one.
+
List item one continued with a second paragraph followed by an
Indented block.
+
.....
$ ls *.sh
$ mv *.sh ~/tmp
......
+
List item continued with a third paragraph.
```

```
2. List item two continued with an open block.

+
--
This paragraph is part of the preceding list item.
a. This list is nested and does not require explicit item continuation.

+
This paragraph is part of the preceding list item.
b. List item b.
```

And here is the equivalent in Markdown:

1. List item one.

List item one continued with a second paragraph followed by an Indented block.

List item continued with a third paragraph.

2. List item two continued with an open block.

This paragraph is part of the preceding list item.

1. This list is nested and does not require explicit item continuation.

This paragraph is part of the preceding list item.

2. List item b.

This paragraph belongs to item two of the outer list.

The AsciiDoc version is, arguably, easier to write. You don't need to worry about indentation. But the Markdown version is much easier to read. The nesting of list items is apparent to the eye in the source, not just in the processed document.

1.3 Why is a spec needed?

John Gruber's <u>canonical description of Markdown's syntax</u> does not specify the syntax unambiguously. Here are some examples of questions it does not answer:

- 1. How much indentation is needed for a sublist? The spec says that continuation paragraphs need to be indented four spaces, but is not fully explicit about sublists. It is natural to think that they, too, must be indented four spaces, but Markdown. pl does not require that. This is hardly a "corner case," and divergences between implementations on this issue often lead to surprises for users in real documents. (See this comment by John Gruber.)
- Is a blank line needed before a block quote or heading? Most implementations do not require the blank line. However, this can lead to unexpected results in

hard-wrapped text, and also to ambiguities in parsing (note that some implementations put the heading inside the blockquote, while others do not). (John Gruber has also spoken in favor of requiring the blank lines.)

3. Is a blank line needed before an indented code block? (Markdown. p1 requires it, but this is not mentioned in the documentation, and some implementations do not require it.)

```
paragraph code?
```

- 4. What is the exact rule for determining when list items get wrapped in tags? Can a list be partially "loose" and partially "tight"? What should we do with a list like this?
 - 1. one
 - 2. two
 - 3. three

Or this?

- 1. one a
- b
- 2. two

(There are some relevant comments by John Gruber here.)

5. Can list markers be indented? Can ordered list markers be right-aligned?

```
8. item 1
9. item 2
10. item 2a
```

6. Is this one list with a thematic break in its second item, or two lists separated by a thematic break?

```
* a
* * * * *
* b
```

- 7. When list markers change from numbers to bullets, do we have two lists or one? (The Markdown syntax description suggests two, but the perl scripts and many other implementations produce one.)
 - 1. fee
 - 2. fie
 - foe
 - fum

8. What are the precedence rules for the markers of inline structure? For example, is the following a valid link, or does the code span take precedence?

```
[a backtick (`)](/url) and [another backtick (`)](/url).
```

9. What are the precedence rules for markers of emphasis and strong emphasis? For example, how should the following be parsed?

```
*foo *bar* baz*
```

10. What are the precedence rules between block-level and inline-level structure? For example, how should the following be parsed?

```
- `a long code span can contain a hyphen like this - and it can screw things up`
```

11. Can list items include section headings? (Markdown. pl does not allow this, but does allow blockquotes to include headings.)

```
- # Heading
```

12. Can list items be empty?

```
* a
*
* b
```

13. Can link references be defined inside block quotes or list items?

```
> Blockquote [foo].
>
> [foo]: /url
```

14. If there are multiple definitions for the same reference, which takes precedence?

```
[foo]: /url1
[foo]: /url2
```

In the absence of a spec, early implementers consulted Markdown. pl to resolve these ambiguities. But Markdown. pl was quite buggy, and gave manifestly bad results in many cases, so it was not a satisfactory replacement for a spec.

Because there is no unambiguous spec, implementations have diverged considerably. As a result, users are often surprised to find that a document that renders one way on one system (say, a GitHub wiki) renders differently on another (say, converting to docbook using pandoc). To make matters worse, because nothing in Markdown counts as a "syntax error," the divergence often isn't discovered right away.

1.4 About this document

This document attempts to specify Markdown syntax unambiguously. It contains many examples with side-by-side Markdown and HTML. These are intended to double as conformance tests. An accompanying script spec_tests.py can be used to run the tests against any Markdown program:

```
python test/spec tests.py --spec spec.txt --program PROGRAM
```

Since this document describes how Markdown is to be parsed into an abstract syntax tree, it would have made sense to use an abstract representation of the syntax tree instead of HTML. But HTML is capable of representing the structural distinctions we need to make, and the choice of HTML for the tests makes it possible to run the tests against an implementation without writing an abstract syntax tree renderer.

This document is generated from a text file, spec. txt, written in Markdown with a small extension for the side-by-side tests. The script tools/makespec.py can be used to convert spec. txt into HTML or CommonMark (which can then be converted into other formats).

In the examples, the \rightarrow character is used to represent tabs.

2 Preliminaries

2.1 Characters and lines

Any sequence of characters is a valid CommonMark document.

A <u>character</u> is a Unicode code point. Although some code points (for example, combining accents) do not correspond to characters in an intuitive sense, all code points count as characters for purposes of this spec.

 \mathbf{A}

This spec does not specify an encoding; it thinks of lines as composed of <u>characters</u> rather than bytes. A conforming parser may be limited to a certain encoding.

A <u>line</u> is a sequence of zero or more <u>characters</u> other than newline (U+000A) or carriage return (U+000D), followed by a <u>line ending</u> or by the end of file.

A <u>line ending</u> is a newline (U+000A), a carriage return (U+000D) not followed by a newline, or a carriage return and a following newline.

A line containing no characters, or a line containing only spaces (U+0020) or tabs (U+0009), is called a **blank line**.

The following definitions of character classes will be used in this spec:

A <u>whitespace character</u> is a space (U+0020), tab (U+0009), newline (U+000A), line tabulation (U+000B), form feed (U+000C), or carriage return (U+000D).

Whitespace is a sequence of one or more whitespace characters.

A <u>Unicode whitespace character</u> is any code point in the Unicode <u>Zs</u> general category, or a tab (U+0009), carriage return (U+000D), newline (U+000A), or form feed (U+000C).

Unicode whitespace is a sequence of one or more **Unicode whitespace** characters.

A **space** is U+0020.

A non-whitespace character is any character that is not a whitespace character.

An <u>ASCII punctuation character</u> is !, ", #, \$, %, &, ', (,), *, +, \, -, \, \ (U+0021–2F), :, ;, \langle , =, \rangle , ?, @ (U+003A–0040), [, \,], ^, _, \) (U+005B–0060), {, ||, ||, ||, or \sim (U+007B–007E).

A <u>punctuation character</u> is an <u>ASCII punctuation character</u> or anything in the general Unicode categories Pc, Pd, Pe, Pf, Pi, Po, or Ps.

2.2 Tabs

Tabs in lines are not expanded to <u>spaces</u>. However, in contexts where whitespace helps to define block structure, tabs behave as if they were replaced by spaces with a tab stop of 4 characters.

Thus, for example, a tab can be used instead of four spaces in an indented code block. (Note, however, that internal tabs are passed through as literal tabs, not expanded to spaces.)

Example 1

```
→foo→baz→→bim </code>foo→baz→→bim
</code>
```

Example 2

```
·→foo→baz→→bim </code>foo→baz→→bim
</code>
```

Example 3

```
···· a→a
···· ὑ→a

ὑ→a
⟨/code⟩⟨/pre⟩
```

In the following example, a continuation paragraph of a list item is indented with a tab; this has exactly the same effect as indentation with four spaces would:

Example 4

Example 5

Normally the > that begins a block quote may be followed optionally by a space, which is not considered part of the content. In the following case > is followed by a tab, which is treated as if it were expanded into three spaces. Since one of these spaces is considered part of the delimiter, foo is considered to be indented six spaces inside the block quote context, so we get an indented code block starting with two spaces.

Example 6

Example 7

```
-→→foo

    ⟨ul⟩
    ⟨pre⟩⟨code⟩ ∘ foo
    ⟨code⟩⟨pre⟩
    ⟨li⟩
    ⟨ul⟩
```

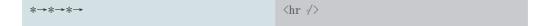
Example 8

Example 9

Example 10

```
#→Foo <h1>Foo</h1>
```

Example 11



2.3 Insecure characters

- حالا حالان. . اد.

For security reasons, the Unicode character U+0000 must be replaced with the REPLACEMENT CHARACTER (U+FFFD).

3 Blocks and inlines

We can think of a document as a sequence of <u>blocks</u>—structural elements like paragraphs, block quotations, lists, headings, rules, and code blocks. Some blocks (like block quotes and list items) contain other blocks; others (like headings and paragraphs) contain <u>inline</u> content—text, links, emphasized text, images, code spans, and so on.

3.1 Precedence

Indicators of block structure always take precedence over indicators of inline structure. So, for example, the following is a list with two items, not a list with one item containing a code span:

Example 12

This means that parsing can proceed in two steps: first, the block structure of the document can be discerned; second, text lines inside paragraphs, headings, and other block constructs can be parsed for inline structure. The second step requires information about link reference definitions that will be available only at the end of the first step. Note that the first step requires processing lines in sequence, but the second can be parallelized, since the inline parsing of one block element does not affect the inline parsing of any other.

3.2 Container blocks and leaf blocks

We can divide blocks into two types: **container blocks**, which can contain other blocks, and **leaf blocks**, which cannot.

4 Leaf blocks

This section describes the different kinds of leaf block that make up a Markdown document.

4.1 Thematic breaks



A line consisting of 0-3 spaces of indentation, followed by a sequence of three or more matching -, -, or - characters, each followed optionally by any number of spaces or tabs, forms a **thematic break**.



Wrong characters:

Example 14

+++	+++

Example 15



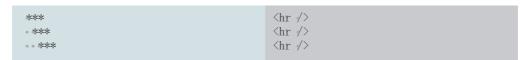
Not enough characters:

Example 16



One to three spaces indent are allowed:

Example 17



Four spaces is too many:

Example 18



Example 19



More than three characters may be used:

Example 20



Spaces are allowed between the characters:

```
Frample 22
```

Example 22

Example 23

Spaces are allowed at the end:

Example 24

```
---- <hr />
```

However, no other characters may occur in the line:

Example 25

It is required that all of the <u>non-whitespace characters</u> be the same. So, this is not a thematic break:

Example 26

Thematic breaks do not need blank lines before or after:

Example 27

Thematic breaks can interrupt a paragraph:

Example 28

If a line of dashes that meets the above conditions for being a thematic break could also be interpreted as the underline of a <u>setext heading</u>, the interpretation as a <u>setext heading</u> takes precedence. Thus, for example, this is a setext heading, not a paragraph followed by a thematic break:

Example 29

When both a thematic break and a list item are possible interpretations of a line, the thematic break takes precedence:

Example 30

If you want a thematic break in a list item, use a different bullet:

Example 31

4.2 ATX headings

An <u>ATX heading</u> consists of a string of characters, parsed as inline content, between an opening sequence of 1–6 unescaped # characters and an optional closing sequence of any number of unescaped # characters. The opening sequence of # characters must be followed by a <u>space</u> or by the end of line. The optional closing sequence of #s must be preceded by a <u>space</u> and may be followed by spaces only. The opening # character may be indented 0-3 spaces. The raw contents of the heading are stripped of leading and trailing spaces before being parsed as inline content. The heading level is equal to the number of # characters in the opening sequence.

Simple headings:

# foo	<h1>foo</h1>
## foo	$\langle h2 \rangle foo \langle /h2 \rangle$
### foo	<h3>foo</h3>
#### f oo	$\langle h4 \rangle foo \langle /h4 \rangle$
##### foo	<h5>foo</h5>
##### foo	<h6>foo</h6>

More than six # characters is not a heading:

Example 33

```
####### foo
```

At least one space is required between the # characters and the heading's contents, unless the heading is empty. Note that many implementations currently do not require the space. However, the space was required by the <u>original ATX implementation</u>, and it helps prevent things like the following from being parsed as headings:

Example 34

This is not a heading, because the first # is escaped:

Example 35

Contents are parsed as inlines:

Example 36

Leading and trailing whitespace is ignored in parsing inline content:

Example 37

One to three spaces indentation are allowed:

Example 38

### foo • ## foo • • # foo	<h3>foo</h3> <h2>foo</h2> <h1>foo</h1>
* # 100	\III/100\/ III/

Four spaces are too much:

Example 39

Example 40



A closing sequence of # characters is optional:

Example 41

It need not be the same length as the opening sequence:

Example 42

Spaces are allowed after the closing sequence:

Example 43

```
### foo ### ••••• <h3>foo</h3>
```

A sequence of # characters with anything but <u>spaces</u> following it is not a closing sequence, but counts as part of the contents of the heading:

Example 44

The closing sequence must be preceded by a space:

Example 45



Backslash-escaped ♯ characters do not count as part of the closing sequence:

Example 46

### foo \###	<h3>foo ###</h3>
## foo #\##	<h2>foo ###</h2>
# foo \#	<h1>foo #</h1>

ATX headings need not be separated from surrounding content by blank lines, and they can interrupt paragraphs:

Example 47

***	<hr/>
## foo	<h2>foo</h2>
****	<hr/>

Example 48

Foo bar	Foo bar
# baz	<h1>baz</h1>
Bar foo	Sar foo

ATX headings can be empty:

Example 49

## •	<h2></h2>
#	<h1></h1>
### ###	<h3></h3>

4.3 Setext headings

A <u>setext heading</u> consists of one or more lines of text, each containing at least one <u>non-whitespace character</u>, with no more than 3 spaces indentation, followed by a <u>setext heading underline</u>. The lines of text must be such that, were they not followed by the setext heading underline, they would be interpreted as a paragraph: they cannot be interpretable as a <u>code fence</u>, <u>ATX heading</u>, <u>block quote</u>, <u>thematic break</u>, <u>list item</u>, or <u>HTML block</u>.

A <u>setext heading underline</u> is a sequence of <u>characters</u> or a sequence of characters, with no more than 3 spaces indentation and any number of trailing spaces. If a line containing a single <u>can be interpreted</u> as an empty <u>list items</u>, it should be interpreted this way and not as a <u>setext heading underline</u>.

The heading is a level 1 heading if characters are used in the <u>setext heading</u> <u>underline</u>, and a level 2 heading if characters are used. The contents of the heading are the result of parsing the preceding lines of text as CommonMark inline content.

In general, a setext heading need not be preceded or followed by a blank line. However, it cannot interrupt a paragraph, so when a setext heading comes after a paragraph, a blank line is needed between them.

Simple examples:

Example 50

```
Foo *bar*
=======

Foo *bar*
-----

Foo *bar*
-----
```

The content of the header may span more than one line:

Example 51



The contents are the result of parsing the headings's raw content as inlines. The heading's raw content is formed by concatenating the lines and removing initial and final whitespace.

Example 52

```
- Foo *bar
baz*→
====
```

The underlining can be any length:

Example 53



The heading content can be indented up to three spaces, and need not line up with the underlining:

Example 54



Four spaces indent is too much:

Example 55

```
***Foo
---
Foo
---
***Code>Foo
---
***Code>
</code>
```

The setext heading underline can be indented up to three spaces, and may have trailing spaces:

Example 56

Four spaces is too much:

Example 57

The setext heading underline cannot contain internal spaces:

Example 58

Trailing spaces in the content line do not cause a line break:

Example 59

Nor does a backslash at the end:

Example 60



Since indicators of block structure take precedence over indicators of inline structure, the following are setext headings:

Example 61

The setext heading underline cannot be a <u>lazy continuation line</u> in a list item or block quote:

Example 62

Example 63

Example 64

A blank line is needed between a paragraph and a following setext heading, since otherwise the paragraph becomes part of the heading's content:

Example 65

But in general a blank line is not required before or after setext headings:

Example 66

Setext headings cannot be empty:

Example 67

```
====
```

Setext heading text lines must not be interpretable as block constructs other than paragraphs. So, the line of dashes in these examples gets interpreted as a thematic break:

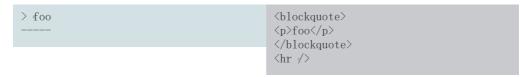
Example 68

Example 69



Example 70

Example 71



If you want a heading with > foo as its literal text, you can use backslash escapes:

Example 72



Compatibility note: Most existing Markdown implementations do not allow the text of setext headings to span multiple lines. But there is no consensus about how to interpret

Foo bar --baz

One can find four different interpretations:

- 1. paragraph "Foo", heading "bar", paragraph "baz"
- 2. paragraph "Foo bar", thematic break, paragraph "baz"
- 3. paragraph "Foo bar baz"
- 4. heading "Foo bar", paragraph "baz"

We find interpretation 4 most natural, and interpretation 4 increases the expressive power of CommonMark, by allowing multiline headings. Authors who want interpretation 1 can put a blank line after the first paragraph:

Example 73



Authors who want interpretation 2 can put blank lines around the thematic break,

Example 74

Foo bar	Foo bar <hr/> baz
baz	

or use a thematic break that cannot count as a setext heading underline, such as

Example 75

Foo bar	Foo bar
* * *	<hr/>
baz	baz

Authors who want interpretation 3 can use backslash escapes:

Example 76

o r	Foo bar
Z	baz

4.4 Indented code blocks

An <u>indented code block</u> is composed of one or more <u>indented chunks</u> separated by blank lines. An <u>indented chunk</u> is a sequence of non-blank lines, each indented four or more spaces. The contents of the code block are the literal contents of the lines, including trailing <u>line endings</u>, minus four spaces of indentation. An indented code block has no <u>info string</u>.

An indented code block cannot interrupt a paragraph, so there must be a blank line between a paragraph and a following indented code block. (A blank line is not needed, however, between a code block and a following paragraph.)

Example 77

```
**** a simple
**** indented code block

'indented code block

</code>
```

If there is any ambiguity between an interpretation of indentation as a code block and as indicating that material belongs to a <u>list item</u>, the list item interpretation takes precedence:

Example 78

Example 79

The contents of a code block are literal text, and do not get parsed as Markdown:

Example 80

Here we have three chunks separated by blank lines:

Example 81

Any initial spaces beyond four will be included in the content, even in interior blank lines:

Example 82

```
chunk1
contact
chunk2
code>chunk1
contact
chunk2
chunk2
</code>
```

An indented code block cannot interrupt a paragraph. (This allows hanging indents and the like.)

Example 83

However, any non-blank line with fewer than four leading spaces ends the code block immediately. So a paragraph may occur immediately after indented code:

Example 84

And indented code can occur immediately before and after other kinds of blocks:

Example 85

The first line can be indented more than four spaces:

Example 86

```
code>
code>

code>
```

Blank lines preceding or following an indented code block are not included in it:

Example 87

```
<code>foo
....foo
.
```

Trailing spaces are included in the code block's content:

Example 88

4.5 Fenced code blocks

A <u>code fence</u> is a sequence of at least three consecutive backtick characters () or tildes (). (Tildes and backticks cannot be mixed.) A <u>fenced code block</u> begins with a code fence, indented no more than three spaces.

The line with the opening code fence may optionally contain some text following the code fence; this is trimmed of leading and trailing whitespace and called the infostring comes after a backtick fence, it may not contain any backtick characters. (The reason for this restriction is that otherwise some inline code would be incorrectly interpreted as the beginning of a fenced code block.)

The content of the code block consists of all subsequent lines, until a closing <u>code fence</u> of the same type as the code block began with (backticks or tildes), and with at least as many backticks or tildes as the opening code fence. If the leading code fence is indented N spaces, then up to N spaces of indentation are removed from each line of the content (if present). (If a content line is not indented, it is preserved unchanged. If it is indented less than N spaces, all of the indentation is removed.)

The closing code fence may be indented up to three spaces, and may be followed only by spaces, which are ignored. If the end of the containing block (or document) is reached and no closing code fence has been found, the code block contains all of the lines after the opening code fence until the end of the containing block (or document). (An alternative spec would require backtracking in the event that a closing code fence is not found. But this makes parsing much less efficient, and there seems to be no real down side to the behavior described here.)

A fenced code block may interrupt a paragraph, and does not require a blank line either before or after.

The content of a code fence is treated as literal text, not parsed as inlines. The first word of the <u>info string</u> is typically used to specify the language of the code sample,

and rendered in the class attribute of the code tag. However, this spec does not mandate any particular treatment of the info string.

Here is a simple example with backticks:

Example 89

```
<code>&lt;
```

With tildes:

Example 90

Fewer than three backticks is not enough:

Example 91

```
foo <code>foo
```

The closing code fence must use the same character as the opening fence:

Example 92

Example 93

```
aaa

</code>
```

The closing code fence must be at least as long as the opening fence:

Example 94

```
aaa </code>
```

Unclosed code blocks are closed by the end of the document (or the enclosing <u>block</u> <u>quote</u> or <u>list item</u>):

Example 96

. . .

```
Example 97

in the second second
```

A code block can have all empty lines as its content:

Example 99

A code block can be empty:

Example 100

```
</code>
```

Fences can be indented. If the opening fence is indented, content lines will have equivalent opening indentation removed, if present:

Example 101

Example 103

Four spaces indentation produces an indented code block:

Example 104

Closing fences may be indented by 0-3 spaces, and their indentation need not match that of the opening fence:

Example 105

Example 106

This is not a closing fence, because it is indented 4 spaces:

Example 107

Code fences (opening and closing) cannot contain internal spaces:

Example 108

```
aaa <code> </code> aaa
```

Example 109

Fenced code blocks can interrupt paragraphs, and can be followed directly by paragraphs, without a blank line between:

Example 110

Other blocks can also occur before and after fenced code blocks without an intervening blank line:

Example 111

An <u>info string</u> can be provided after the opening code fence. Although this spec doesn't mandate any particular treatment of the info string, the first word is typically used to specify the language of the code block. In HTML output, the language is normally indicated by adding a class to the <u>code</u> element consisting of <u>language</u>-followed by the language name.

Example 112

```
ruby
def foo(x)
return 3
end
end
foo(x)
foo(x)
return 3
end
foo(x)
return 3
end
foo(x)
return 3
end
foo(x)
return 3
```

Example 113

```
code class="language-;"></code>
class="language-;"></code>
code class="language-;"></code clas
```

Info strings for backtick code blocks cannot contain backticks:

Example 115

```
foo \( \frac{p}{\code} \aa \land{\code} \\ foo \( \frac{p}{\code} \)
```

Info strings for tilde code blocks can contain backticks and tildes:

Example 116

```
code class="language-aa">foo
foo
code>
```

Closing code fences cannot have info strings:

Example 117

4.6 HTML blocks

An <u>HTML block</u> is a group of lines that is treated as raw HTML (and will not be escaped in HTML output).

A

There are seven kinds of <a href="https://http

1. **Start condition:** line begins with the string \(\script \), \(\script \), or \(\style \) (case-insensitive), followed by whitespace, the string \(\script \), \(\s

- 2. **Start condition:** line begins with the string <!--. **End condition:** line contains the string -->.
- 3. **Start condition:** line begins with the string <?. **End condition:** line contains the string ?>.
- Start condition: line begins with the string <! followed by an uppercase ASCII letter.

End condition: line contains the character >.

- 5. **Start condition:** line begins with the string <! [CDATA]. **End condition:** line contains the string]]>.
- 6. Start condition: line begins the string < or </ followed by one of the strings (case-insensitive) address, article, aside, base, basefont, blockquote, body, caption, center, col, colgroup, dd, details, dialog, dir, div, dl, dt, fieldset, figcaption, figure, footer, form, frame, frameset, h1, h2, h3, h4, h5, h6, head, header, hr, html, iframe, legend, li, link, main, menu, menuitem, nav, noframes, ol, optgroup, option, p, param, section, source, summary, table, tbody, td, tfoot, th, thead, title, tr, track, ul, followed by whitespace, the end of the line, the string >, or the string />.

End condition: line is followed by a blank line.

7. **Start condition:** line begins with a complete <u>open tag</u> (with any <u>tag name</u> other than <u>script</u>, <u>style</u>, or <u>pre</u>) or a complete <u>closing tag</u>, followed only by <u>whitespace</u> or the end of the line.

End condition: line is followed by a blank line.

HTML blocks continue until they are closed by their appropriate <u>end condition</u>, or the last line of the document or other <u>container block</u>. This means any HTML **within an HTML block** that might otherwise be recognised as a start condition will be ignored by the parser and passed through as-is, without changing the parser's state.

For instance, within a HTML block started by will not affect the parser state; as the HTML block was started in by start condition 6, it will end at any blank line. This can be surprising:

Example 118

```
    \langle tr \langle tr \langle td \rangle tr \langle tr \langle td \rangle tr \langle t
```

In this case, the HTML block is terminated by the newline — the ** ${\rm He}10**$ text remains verbatim — and regular parsing resumes, with a paragraph, emphasised world and inline and block HTML following.

All types of <u>HTML blocks</u> except type 7 may interrupt a paragraph. Blocks of type 7 may not interrupt a paragraph. (This restriction is intended to prevent unwanted interpretation of long tags inside a wrapped paragraph as starting HTML blocks.)

Some simple examples follow. Here are some basic HTML blocks of type 6:

```
• 
                           • 
•••
                           ••• 
····hi
                           ····hi
•••
                           •••
•
                           •
\langle p \rangleokay.\langle /p \rangle
okay.
```

Example 120

A block can also start with a closing tag:

Example 121

foo	*foo*

Here we have two HTML blocks with a Markdown paragraph between them:

Example 122

The tag on the first line can be partial, as long as it is split where there would be whitespace:

Example 123

Example 124

An open tag need not be closed:

Example 125

<div></div>	<div></div>
foo	*foo*
	bar
har	

A partial tag need not even be completed (garbage in, garbage out):

Example 126

Example 127

<div class<="" th=""><th><div class<="" th=""></div></th></div>	<div class<="" th=""></div>
foo	foo

The initial tag doesn't even need to be a valid tag, as long as it starts like one:

Example 128

```
<div *???-&&-<---
*foo*</pre>
<div *???-&&&-<---
*foo*</pre>
```

In type 6 blocks, the initial tag need not be on a line by itself:

Example 129

<div>*foo*</div>	<div>*foo*</div>

Example 130

foo	foo
//tu//ti//table/	\/ tu/\/ ti/\/ table/

Everything until the next blank line or end of document gets included in the HTML block. So, in the following example, what looks like a Markdown code block is actually part of the HTML block, which continues until a blank line or the end of the document is reached:

Example 131

To start an <u>HTML block</u> with a tag that is *not* in the list of block-level tags in (6), you must put the tag by itself on the first line (and it must be complete):

Example 132

	
bar	*bar*
⟨/a⟩	

In type 7 blocks, the tag name can be anything:

Example 133

<warning></warning>	<warning></warning>
bar	*bar*

Example 134

Example 135

bar	*bar*

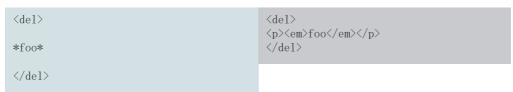
These rules are designed to allow us to work with tags that can function as either block-level or inline-level tags. The $\langle de1 \rangle$ tag is a nice example. We can surround content with $\langle de1 \rangle$ tags in three different ways. In this case, we get a raw HTML block, because the $\langle de1 \rangle$ tag is on a line by itself:

Example 136

	
foo	*foo*

In this case, we get a raw HTML block that just includes the $\langle de1 \rangle$ tag (because it ends with the following blank line). So the contents get interpreted as CommonMark:

Example 137



Finally, in this case, the $\langle de1 \rangle$ tags are interpreted as <u>raw HTML</u> inside the CommonMark paragraph. (Because the tag is not on a line by itself, we get inline HTML rather than an HTML block.)

Example 138

foo	$\label{eq:condition} $$\langle p \rangle \langle del \rangle \langle em \rangle foo \langle em \rangle \langle del \rangle \langle p \rangle $$$
------------------	---

HTML tags designed to contain literal content (script, style, pre), comments, processing instructions, and declarations are treated somewhat differently. Instead of ending at the first blank line, these blocks end at the first line containing a corresponding end tag. As a result, these blocks can contain blank lines:

A pre tag (type 1):

Example 139

A script tag (type 1):

Example 140

```
<script type="text/javascript">
// JavaScript example

document.getElementById("demo").innerHT
ML = "Hello JavaScript!";
</script>
okay

<script type="text/javascript">
// JavaScript example

document.getElementById("demo").innerHT
ML = "Hello JavaScript!";
</script>
okay
```

A style tag (type 1):

Example 141

If there is no matching end tag, the block will end at the end of the document (or the enclosing <u>block quote</u> or <u>list item</u>):

Example 142

Example 143

> <div> foo</div>	 div>
bar	foo bar
	\p/ bar \/ p/

```
 \begin{array}{c} - \ \langle \operatorname{div} \rangle \\ - \ foo \end{array}
```

The end tag can occur on the same line as the start tag:

Example 145

```
 \begin{array}{ll} \langle style \rangle p \{color:red;\} \langle /style \rangle & \langle style \rangle p \{color:red;\} \langle /style \rangle \\ *foo* & \langle p \rangle \langle em \rangle foo \langle /em \rangle \langle /p \rangle \\ \end{array}
```

Example 146

Note that anything on the last line after the end tag will be included in the <u>HTML block</u>:

Example 147

A comment (type 2):

Example 148

```
      <!-- Foo</td>
      <!-- Foo</td>

      bar
      bar

      -- baz
      -->

      okay
      okay
```

A processing instruction (type 3):

Example 149

```
</php

- echo '>';

?>
okay

<p
```

A declaration (type 4):

Example 150

```
<!DOCTYPE html> <!DOCTYPE html>
```

CDATA (type 5):

Example 151

```
<![CDATA[
                                         <![CDATA[
function matchwo(a, b)
                                         function matchwo(a,b)
•if (a < b && a < 0) then {
                                          if (a < b && a < 0) then {
return 1;
                                          return 1;
•} else {
                                          •} else {
••• return 0;
                                          · · · return 0;
                                          • }
• •}
]]>
                                         ]]>
                                         okay
okay
```

The opening tag can be indented 1-3 spaces, but not 4:

Example 152

Example 153

```
-<div>
-<div>
<code>&lt;div&gt;
</code>
```

An HTML block of types 1–6 can interrupt a paragraph, and need not be preceded by a blank line.

Example 154

Foo	Foo
<div></div>	<div></div>
bar	bar

However, a following blank line is needed, except at the end of a document, and except for blocks of types 1–5, <u>above</u>:

Example 155

<div></div>	<div></div>
bar	bar
foo	*foo*

HTML blocks of type 7 cannot interrupt a paragraph:

Example 156

This rule differs from John Gruber's original Markdown syntax specification, which says:

The only restrictions are that block-level HTML elements — e.g. $\langle \text{div} \rangle$, $\langle \text{table} \rangle$, $\langle \text{pre} \rangle$, $\langle \text{p} \rangle$, etc. — must be separated from surrounding content by blank lines, and the start and end tags of the block should not be indented with tabs or spaces.

In some ways Gruber's rule is more restrictive than the one given here:

- It requires that an HTML block be preceded by a blank line.
- It does not allow the start tag to be indented.
- It requires a matching end tag, which it also does not allow to be indented.

Most Markdown implementations (including some of Gruber's own) do not respect all of these restrictions.

There is one respect, however, in which Gruber's rule is more liberal than the one given here, since it allows blank lines to occur inside an HTML block. There are two reasons for disallowing them here. First, it removes the need to parse balanced tags, which is expensive and can require backtracking from the end of the document if no matching end tag is found. Second, it provides a very simple and flexible way of including Markdown content inside HTML tags: simply separate the Markdown from the HTML using blank lines:

Compare:

Example 157

```
<div>
*Emphasized* text.
</div>
*Emphasized* text.
</div>
*Emphasized* text.
</div>
```

Some Markdown implementations have adopted a convention of interpreting content inside tags as text if the open tag has the attribute markdown=1. The rule given above seems a simpler and more elegant way of achieving the same expressive power, which is also much simpler to parse.

The main potential drawback is that one can no longer paste HTML blocks into Markdown documents with 100% reliability. However, *in most cases* this will work fine, because the blank lines in HTML are usually followed by HTML block tags. For example:

Example 159

```
    Hi
```

There are problems, however, if the inner tags are indented *and* separated by spaces, as then they will be interpreted as an indented code block:

Example 160

Fortunately, blank lines are usually not necessary and can be deleted. The exception is inside tags, but as described above, raw HTML blocks starting with can contain blank lines.

4.7 Link reference definitions

A <u>link reference definition</u> consists of a <u>link label</u>, indented up to three spaces, followed by a colon (:), optional <u>whitespace</u> (including up to one <u>line ending</u>), a <u>link destination</u>, optional <u>whitespace</u> (including up to one <u>line ending</u>), and an optional <u>link title</u>, which if it is present must be separated from the <u>link destination</u> by <u>whitespace</u>. No further <u>non-whitespace characters</u> may occur on the line.

A <u>link reference definition</u> does not correspond to a structural element of a document. Instead, it defines a label which can be used in <u>reference links</u> and reference-style <u>images</u> elsewhere in the document. <u>Link reference definitions</u> can come either before or after the links that use them.

Example 161

Example 162

Example 163

Example 164

```
[Foo bar]:
<my url>
'title'

[Foo bar]

<a href="my%20url" title="title">Foo o
bar</a>
```

The title may extend over multiple lines:

Example 165

```
[foo]: /url '
title
title
line1
line2
'

[foo]
```

However, it may not contain a blank line:

Example 166

The title may be omitted:

Example 167

The link destination may not be omitted:

Example 168

However, an empty link destination may be specified using angle brackets:

Example 169

The title must be separated from the link destination by whitespace:

Example 170

```
[foo]: \( \data \rangle \text{(baz)} \) \( \langle \p \rangle \text{[foo]: \( \data \rangle \text{(baz)} \langle \p \rangle \text{[foo]} \) \( \langle \p \rangle \text{[foo] \( \data \rangle \text{[foo]} \langle \p \rangle \text{[foo]} \)
```

Both title and destination can contain backslash escapes and literal backslashes:

Example 171

```
[foo]: \dirplan \*baz \"foo\"bar\baz" \qquad \text{title="foo"bar\baz" \qquad \text{title="foo"bar\baz"} \qquad \text{foo} \qquad \text{p} \qquad \text{title="foo"bar\baz"} \qquad \text{foo} \qquad \qqquad \qqquad \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqq \qqqq \qqqqq \qqqq \qqq \qqqq \q
```

A link can come before its corresponding definition:

Example 172

If there are several matching definitions, the first one takes precedence:

Example 173

```
[foo]: first [foo]: second
```

As noted in the section on <u>Links</u>, matching of labels is case-insensitive (see <u>matches</u>).

Example 174

Example 175

Here is a link reference definition with no corresponding link. It contributes nothing to the document.

Example 176

```
[foo]: /url
```

Here is another one:

Example 177

```
[
foo
]: \forall url
bar
```

This is not a link reference definition, because there are <u>non-whitespace characters</u> after the title:

Example 178

This is a link reference definition, but it has no title:

Example 179

This is not a link reference definition, because it is indented four spaces:

Example 180

This is not a link reference definition, because it occurs inside a code block:

Example 181

A link reference definition cannot interrupt a paragraph.

Example 182

However, it can directly follow other block elements, such as headings and thematic breaks, and it need not be followed by a blank line.

Example 183

```
[foo]: /url
bar
===
[foo]
```

Example 185

Several <u>link reference definitions</u> can occur one after another, without intervening blank lines.

Example 186

<u>Link reference definitions</u> can occur inside block containers, like lists and block quotations. They affect the entire document, not just the container in which they are defined:

Example 187

Whether something is a <u>link reference definition</u> is independent of whether the link reference it defines is used in the document. Thus, for example, the following document contains just a link reference definition, and no visible content:

Example 188

```
[foo]: √url
```

A

4.8 Paragraphs

https://github.github.com/gfm/ 42/138

A sequence of non-blank lines that cannot be interpreted as other kinds of blocks forms a **paragraph**. The contents of the paragraph are the result of parsing the paragraph's raw content as inlines. The paragraph's raw content is formed by concatenating the lines and removing initial and final **whitespace**.

A simple example with two paragraphs:

Example 189

Paragraphs can contain multiple lines, but no blank lines:

Example 190

aaa bbb	aaa bbb
ccc	ccc ddd
ddd	

Multiple blank lines between paragraph have no effect:

Example 191

Leading spaces are skipped:

Example 192



Lines after the first may be indented any amount, since indented code blocks cannot interrupt paragraphs.

Example 193



However, the first line may be indented at most three spaces, or an indented code block will be triggered:

aaa bbb	aaa bbb
Example 195	
•••• aaa bbb	<pre><code>aaa </code></pre>

Final spaces are stripped before inline parsing, so a paragraph that ends with two or more spaces will not end with a <u>hard line break</u>:

 $\langle p \rangle bbb \langle /p \rangle$

Example 196

4.9 Blank lines

<u>Blank lines</u> between block-level elements are ignored, except for the role they play in determining whether a <u>list</u> is <u>tight</u> or <u>loose</u>.

Blank lines at the beginning and end of the document are also ignored.

Example 197

4.10 Tables (extension)

GFM enables the table extension, where an additional leaf block type is available.

A <u>table</u> is an arrangement of data with rows and columns, consisting of a single header row, a <u>delimiter row</u> separating the header from the data, and zero or more data rows.

Each row consists of cells containing arbitrary text, in which <u>inlines</u> are parsed, separated by pipes (|). A leading and trailing pipe is also recommended for clarity of reading, and if there's otherwise parsing ambiguity. Spaces between pipes and cell content are trimmed. Block-level elements cannot be inserted in a table.

The <u>delimiter row</u> consists of cells whose only content are hyphens (-), and optionally, a leading or trailing colon (:), or both, to indicate left, right, or center alignment respectively.

```
foo bar
                                            <thead>
baz bim
                                            \langle \mathrm{tr} \rangle
                                            \langle th \rangle foo \langle /th \rangle
                                            bar
                                            </thead>
                                            >
                                            baz
                                            \langle td \rangle bim \langle /td \rangle
```

Cells in one column don't need to match length, though it's easier to read if they are. Likewise, use of leading and trailing pipes may be inconsistent:

Example 199

```
abc defghi
                           :-: • -----
                           <thead>
bar baz
                           \langle tr \rangle
                           abc
                           defghi
                           </thead>
                           bar
                           \begin{tabular}{ll} $$ \dign="right">baz\\ \end{tabular}
```

Include a pipe in a cell's content by escaping it, including inside other inline spans:

Example 200

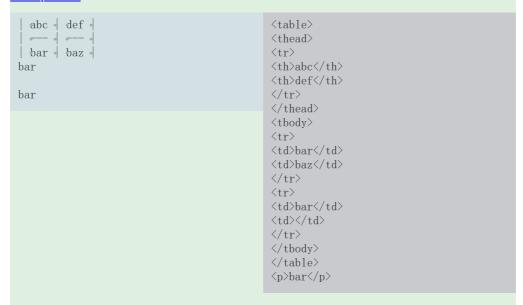
```
f\|00 ••|
                                    <thead>
 b \\| az |
                                    >
| b **\|** im |
                                    \langle th \rangle f | oo \langle /th \rangle
                                    </thead>
                                    \langle tr \rangle
                                    b <code <</d> az
                                    >
                                    b <strong>|</strong> im
```

The table is broken at the first empty line, or beginning of another block-level structure:

Example 201

```
abc | def |
                                                          <thead>
| bar | baz |
                                                          \langle tr \rangle
                                                          \langle th \rangle abc \langle /th \rangle
> bar
                                                          \  \  \langle th\rangle def <\!/th\rangle \\
                                                          </thead>
                                                         >
                                                          \langle td \rangle bar \langle /td \rangle
                                                          baz
                                                          \langle / tr \rangle
                                                         <blookquote>
                                                         \bar
                                                          </blockquote>
```

Example 202



The header row must match the <u>delimiter row</u> in the number of cells. If not, a table will not be recognized:

Example 203

The remainder of the table's rows may vary in the number of cells. If there are a number of cells fewer than the number of cells in the header row, empty cells are inserted. If there are greater, the excess is ignored:

Example 204

```
abc def
                                     <thead>
bar •
                                     >
bar | baz | boo |
                                     abc
                                     \langle th \rangle def \langle /th \rangle
                                     </thead>
                                     \langle tr \rangle
                                     bar
                                     \langle td \rangle \langle /td \rangle
                                     bar
                                     baz
```

If there are no rows in the body, no is generated in HTML output:

Example 205

5 Container blocks

A <u>container block</u> is a block that has other blocks as its contents. There are two basic kinds of container blocks: <u>block quotes</u> and <u>list items</u>. <u>Lists</u> are meta-containers for <u>list items</u>.

We define the syntax for container blocks recursively. The general form of the definition is:

If X is a sequence of blocks, then the result of transforming X in suchand-such a way is a container of type Y with these blocks as its content.

So, we explain what counts as a block quote or list item by explaining how these can be *generated* from their contents. This should suffice to define the syntax, although it does not give a recipe for *parsing* these constructions. (A recipe is provided below in the section entitled <u>A parsing strategy</u>.)

5.1 Block quotes

A **block quote marker** consists of 0-3 spaces of initial indent, plus (a) the character together with a following space, or (b) a single character to not followed by a space.

The following rules define block quotes:

- 1. **Basic case.** If a string of lines *Ls* constitute a sequence of blocks *Bs*, then the result of prepending a <u>block quote marker</u> to the beginning of each line in *Ls* is a <u>block quote</u> containing *Bs*.
- 2. Laziness. If a string of lines Ls constitute a <u>block quote</u> with contents Bs, then the result of deleting the initial <u>block quote marker</u> from one or more lines in which the next <u>non-whitespace character</u> after the <u>block quote marker</u> is <u>paragraph continuation text</u> is a block quote with Bs as its content. <u>Paragraph continuation text</u> is text that will be parsed as part of the content of a paragraph, but does not occur at the beginning of the paragraph.
- 3. **Consecutiveness.** A document cannot contain two <u>block quotes</u> in a row unless there is a <u>blank line</u> between them.

Nothing else counts as a block quote.

Here is a simple example:

Example 206

The spaces after the > characters can be omitted:

Example 207

The characters can be indented 1-3 spaces:

Example 208

Four spaces gives us a code block:

The Laziness clause allows us to omit the before paragraph continuation text:

Example 210

A block quote can contain some lazy and some non-lazy continuation lines:

Example 211

```
> bar
baz
> foo

foo

baz
coolonger
co
```

Laziness only applies to lines that would have been continuations of paragraphs had they been prepended with <u>block quote markers</u>. For example, the ∑ cannot be omitted in the second line of

```
> foo
```

without changing the meaning:

Example 212

Similarly, if we omit the > in the second line of

```
> - foo
> - bar
```

then the block quote ends after the first line:

Example 213

For the same reason, we can't omit the $\ \ \ \ \$ in front of subsequent lines of an indented or fenced code block:

Example 214

Example 215

Note that in the following case, we have a <u>lazy continuation line</u>:

Example 216

To see why, note that in

```
> foo
> - bar
```

the bar is indented too far to start a list, and can't be an indented code block because indented code blocks cannot interrupt paragraphs, so it is paragraph continuation text.

A block quote can be empty:

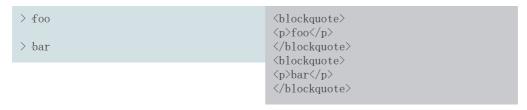
Example 217

A block quote can have initial or final blank lines:

Example 219

A blank line always separates block quotes:

Example 220



(Most current Markdown implementations, including John Gruber's original Markdown. p1, will parse this example as a single block quote with two paragraphs. But it seems better to allow the author to decide whether two block quotes or one are wanted.)

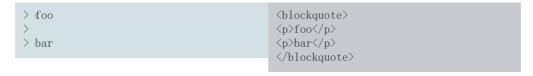
Consecutiveness means that if we put these block quotes together, we get a single block quote:

Example 221



To get a block quote with two paragraphs, use:

Example 222



Block quotes can interrupt paragraphs:

Example 223

In general, blank lines are not needed before or after block quotes:

Example 224

However, because of laziness, a blank line is needed between a block quote and a following paragraph:

Example 225

Example 226

Example 227

It is a consequence of the Laziness rule that any number of initial >s may be omitted on a continuation line of a nested block quote:

Example 228

When including an indented code block in a block quote, remember that the <u>block</u> <u>quote marker</u> includes both the \geq and a following space. So *five spaces* are needed after the \geq :

Example 230

A

5.2 List items

A list marker is a bullet list marker or an ordered list marker.

A **bullet list marker** is a -, +, or * character.

An <u>ordered list marker</u> is a sequence of 1–9 arabic digits (0-9), followed by either a character or a) character. (The reason for the length limit is that with 10 digits we start seeing integer overflows in some browsers.)

The following rules define list items:

1. **Basic case.** If a sequence of lines Ls constitute a sequence of blocks Bs starting with a <u>non-whitespace character</u>, and M is a list marker of width W followed by $1 \le N \le 4$ spaces, then the result of prepending M and the following spaces to the first line of Ls, and indenting subsequent lines of Ls by W+N spaces, is a list item with Bs as its contents. The type of the list item (bullet or ordered) is determined by the type of its list marker. If the list item is ordered, then it is also assigned a start number, based on the ordered list marker.

Exceptions:

- 1. When the first list item in a <u>list</u> interrupts a paragraph—that is, when it starts on a line that would otherwise count as <u>paragraph continuation</u> <u>text</u>—then (a) the lines *Ls* must not begin with a blank line, and (b) if the list item is ordered, the start number must be 1.
- 2. If any line is a thematic break then that line is not a list item.

For example, let Ls be the lines

Example 231

And let M be the marker 1, and N = 2. Then rule #1 says that the following is an ordered list item with start number 1, and the same contents as Ls:

Example 232

```
1. • A paragraph
••••with two lines.

(li)
(p>A paragraph
with two lines. 
(pre><code>indented code
(/code>
(blockquote)
(p>A block quote. 
(blockquote)
(/blockquote)
(/li)
(/ol>
```

The most important thing to notice is that the position of the text after the list marker determines how much indentation is needed in subsequent blocks in the list item. If the list marker takes up two spaces, and there are three spaces between the list marker and the next <u>non-whitespace character</u>, then blocks must be indented five spaces in order to fall under the list item.

Here are some examples showing how far content must be indented to be put under the list item:

Example 233

```
- one  \begin{array}{c} \langle ul \rangle \\ \langle li \rangle one \langle /li \rangle \\ \langle vl \rangle \\ \langle vl \rangle \\ \langle p \rangle two \langle /p \rangle \\ \end{array}
```

Example 234

Example 235

It is tempting to think of this in terms of columns: the continuation blocks must be indented at least to the column of the first <u>non-whitespace character</u> after the list marker. However, that is not quite right. The spaces after the list marker determine how much relative indentation is needed. Which column this indentation reaches will depend on how the list item is embedded in other constructions, as shown by this example:

Example 237

Here two occurs in the same column as the list marker 1., but is actually contained in the list item, because there is sufficient indentation after the last containing blockquote marker.

The converse is also possible. In the following example, the word two occurs far to the right of the initial text of the list item, one, but it is not considered part of the list item, because it is not indented far enough past the blockquote marker:

Example 238

Note that at least one space is needed between the list marker and any following content, so these are not list items:

Example 239

A list item may contain blocks that are separated by more than one blank line.

Example 240

A list item may contain any kind of block:

Example 241

```
1. • foo
                                           <01>
                                           <1i>>
....
                                           \langle p \rangle foo \langle /p \rangle
• • • bar
                                           <code>bar
                                           </code>
                                           baz
· · · baz
                                           <blookquote>
                                           bam
                                           </blockquote>
••• bam
```

A list item that contains an indented code block will preserve empty lines within the code block verbatim.

Example 242

Note that ordered list start numbers must be nine digits or less:

Example 244

1234567890. not ok	1234567890. not ok
--------------------	--------------------

A start number may begin with 0s:

Example 245

Example 246

A start number may not be negative:

Example 247

```
-1. not ok -1. not ok
```

2. **Item starting with indented code.** If a sequence of lines *Ls* constitute a sequence of blocks *Bs* starting with an indented code block, and *M* is a list marker of width *W* followed by one space, then the result of prepending *M* and the following space to the first line of *Ls*, and indenting subsequent lines of *Ls* by *W* + 1 spaces, is a list item with *Bs* as its contents. If a line is empty, then it need not be indented. The type of the list item (bullet or ordered) is determined by the type of its list marker. If the list item is ordered, then it is also assigned a start number, based on the ordered list marker.

An indented code block will have to be indented four spaces beyond the edge of the region where text will be included in the list item. In the following case that is 6 spaces:

Example 248

And in this case it is 11 spaces:

Example 249

If the *first* block in the list item is an indented code block, then by rule #2, the contents must be indented *one* space after the list marker:

Example 250

Example 251

Note that an additional space indent is interpreted as space inside the code block:

Example 252

Note that rules #1 and #2 only apply to two cases: (a) cases in which the lines to be included in a list item begin with a non-whitespace character, and (b) cases in which they begin with an indented code block. In a case like the following, where the first block begins with a three-space indent, the rules do not allow us to form a list item by indenting the whole thing and prepending a list marker:

Example 254

This is not a significant restriction, because when a block begins with 1-3 spaces indent, the indentation can always be removed without a change in interpretation, allowing rule #1 to be applied. So, in the above case:

Example 255

3. **Item starting with a blank line.** If a sequence of lines *Ls* starting with a single blank line constitute a (possibly empty) sequence of blocks *Bs*, not separated from each other by more than one blank line, and *M* is a list marker of width *W*, then the result of prepending *M* to the first line of *Ls*, and indenting subsequent lines of *Ls* by *W* + 1 spaces, is a list item with *Bs* as its contents. If a line is empty, then it need not be indented. The type of the list item (bullet or ordered) is determined by the type of its list marker. If the list item is ordered, then it is also assigned a start number, based on the ordered list marker.

Here are some list items that start with a blank line but are not empty:

Example 256

When the list item starts with a blank line, the number of spaces following the list marker doesn't change the required indentation:

```
- • • • (ul)
• foo (li) foo (li) (/ul)
```

A list item can begin with at most one blank line. In the following example, foo is not part of the list item:

Example 258

Here is an empty bullet list item:

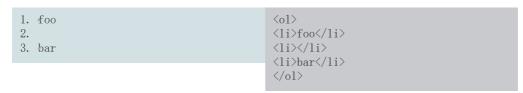
Example 259

It does not matter whether there are spaces following the list marker:

Example 260

Here is an empty ordered list item:

Example 261



A list may start or end with an empty list item:

Example 262

```
 \begin{array}{c} * \\ & \langle ul \rangle \\ & \langle li \rangle \langle /li \rangle \\ & \langle /ul \rangle \end{array}
```

However, an empty list item cannot interrupt a paragraph:

Example 263

4. **Indentation.** If a sequence of lines *Ls* constitutes a list item according to rule #1, #2, or #3, then the result of indenting each line of *Ls* by 1-3 spaces (the same for each line) also constitutes a list item with the same contents and attributes. If a line is empty, then it need not be indented.

Indented one space:

Example 264

Indented two spaces:

Example 265

Indented three spaces:

Example 266

Four spaces indent gives a code block:

Example 267

5. Laziness. If a string of lines Ls constitute a <u>list item</u> with contents Bs, then the result of deleting some or all of the indentation from one or more lines in which the next <u>non-whitespace character</u> after the indentation is <u>paragraph</u> <u>continuation text</u> is a list item with the same contents and attributes. The unindented lines are called <u>lazy continuation lines</u>.

Here is an example with <u>lazy continuation lines</u>:

Example 268

```
***. **A paragraph
with two lines.

***operation*
**outhouse indented code

**outhouse inde
```

Indentation can be partially deleted:

Example 269

```
•1. •A paragraph
•••with two lines.

(ol)
(li>A paragraph
with two lines. 
(/ol>
```

These examples show how laziness can work in nested structures:

Example 270

Example 271

6. **That's all.** Nothing that is not counted as a list item by rules #1–5 counts as a <u>list item</u>.

The rules for sublists follow from the general rules <u>above</u>. A sublist must be indented the same number of spaces a paragraph would need to be in order to be included in the list item.

So, in this case we need two spaces indent:

Example 272

```
- foo
                                      <u1>
• - bar
                                      <1i>foo
••• baz
                                      ul>
•••• boo
                                      <1i>bar
                                      <u1>
                                       <1i>baz
                                      ul>
                                      \langle 1i \rangle boo \langle /1i \rangle
```

One is not enough:

Example 273

Here we need four, because the list marker is wider:

Example 274

Three is not enough:

Example 275

A list may be the first block in a list item:

Example 276

Example 277

```
1. = 2. foo

| Col | Col
```

A list item can contain a heading:

Example 278

5.2.1 Motivation

John Gruber's Markdown spec says the following about list items:

- 1. "List markers typically start at the left margin, but may be indented by up to three spaces. List markers must be followed by one or more spaces or a tab."
- 2. "To make lists look nice, you can wrap items with hanging indents.... But if you don't want to, you don't have to."
- 3. "List items may consist of multiple paragraphs. Each subsequent paragraph in a list item must be indented by either 4 spaces or one tab."
- 4. "It looks nice if you indent every line of the subsequent paragraphs, but here again, Markdown will allow you to be lazy."
- 5. "To put a blockquote within a list item, the blockquote's ≥ delimiters need to be indented."
- To put a code block within a list item, the code block needs to be indented twice — 8 spaces or two tabs."

These rules specify that a paragraph under a list item must be indented four spaces (presumably, from the left margin, rather than the start of the list marker, but this is not said), and that code under a list item must be indented eight spaces instead of the usual four. They also say that a block quote must be indented, but not by how much; however, the example given has four spaces indentation. Although nothing is said about other kinds of block-level content, it is certainly reasonable to infer that *all* block elements under a list item, including other lists, must be indented four spaces. This principle has been called the *four-space rule*.

The four-space rule is clear and principled, and if the reference implementation Markdown. pl had followed it, it probably would have become the standard. However, Markdown. pl allowed paragraphs and sublists to start with only two spaces indentation, at least on the outer level. Worse, its behavior was inconsistent: a sublist of an outer-level list needed two spaces indentation, but a sublist of this sublist needed three spaces. It is not surprising, then, that different implementations of Markdown have developed very different rules for determining what comes under a list item. (Pandoc and python-Markdown, for example, stuck with Gruber's syntax description and the four-space rule, while discount, redcarpet, marked, PHP Markdown, and others followed Markdown. pl's behavior more closely.)

Unfortunately, given the divergences between implementations, there is no way to give a spec for list items that will be guaranteed not to break any existing documents. However, the spec given here should correctly handle lists formatted with either the four-space rule or the more forgiving Markdown. pl behavior, provided they are laid out in a way that is natural for a human to read.

The strategy here is to let the width and indentation of the list marker determine the indentation necessary for blocks to fall under the list item, rather than having a fixed and arbitrary number. The writer can think of the body of the list item as a unit which gets indented to the right enough to fit the list marker (and any indentation on the list marker). (The laziness rule, #5, then allows continuation lines to be unindented if needed.)

This rule is superior, we claim, to any rule requiring a fixed level of indentation from the margin. The four-space rule is clear but unnatural. It is quite unintuitive that

```
- foo
bar
- baz
```

should be parsed as two lists with an intervening paragraph,

```
        foo
        ful>

            pbar

            distant
```

as the four-space rule demands, rather than a single list,

The choice of four spaces is arbitrary. It can be learned, but it is not likely to be guessed, and it trips up beginners regularly.

Would it help to adopt a two-space rule? The problem is that such a rule, together with the rule allowing 1–3 spaces indentation of the initial list marker, allows text that is indented *less than* the original list marker to be included in the list item. For example, Markdown, pl parses

```
- one
```

as a single list item, with two a continuation paragraph:

```
    one
```

and similarly

```
> - one
>
> two
```

as

```
<br/>
```

This is extremely unintuitive.

Rather than requiring a fixed indent from the margin, we could require a fixed indent (say, two spaces, or even one space) from the list marker (which may itself be indented). This proposal would remove the last anomaly discussed. Unlike the spec presented above, it would count the following as a list item with a subparagraph, even though the paragraph bar is not indented as far as the first paragraph foo:

```
10. foo
```

Arguably this text does read like a list item with bar as a subparagraph, which may count in favor of the proposal. However, on this proposal indented code would have to be indented six spaces after the list marker. And this would break a lot of existing Markdown, which has the pattern:

1. foo

indented code

where the code is indented eight spaces. The spec above, by contrast, will parse this text as expected, since the code block's indentation is measured from the beginning of foo.

The one case that needs special treatment is a list item that *starts* with indented code. How much indentation is required in that case, since we don't have a "first paragraph" to measure from? Rule #2 simply stipulates that in such cases, we require one space indentation from the list marker (and then the normal four spaces for the indented code). This will match the four-space rule in cases where the list marker plus its initial indentation takes four spaces (a common case), but diverge in other cases.

5.3 Task list items (extension)

GFM enables the tasklist extension, where an additional processing step is performed on <u>list items</u>.

A <u>task list item</u> is a <u>list item</u> where the first block in it is a paragraph which begins with a <u>task list item marker</u> and at least one whitespace character before any other content.

A <u>task list item marker</u> consists of an optional number of spaces, a left bracket ([), either a whitespace character or the letter \bar{x} in either lowercase or uppercase, and then a right bracket (]).

When rendered, the <u>task list item marker</u> is replaced with a semantic checkbox element; in an HTML output, this would be an <input type="checkbox"> element.

If the character between the brackets is a whitespace character, the checkbox is unchecked. Otherwise, the checkbox is checked.

This spec does not define how the checkbox elements are interacted with: in practice, implementors are free to render the checkboxes as disabled or inmutable elements, or they may dynamically handle dynamic interactions (i.e. checking, unchecking) in the final rendered document.

Example 279

Task lists can be arbitrarily nested:

Example 280

5.4 Lists

A <u>list</u> is a sequence of one or more list items <u>of the same type</u>. The list items may be separated by any number of blank lines.

Two list items are of the same type if they begin with a <u>list marker</u> of the same type. Two list markers are of the same type if (a) they are bullet list markers using the same character (-, +, or *) or (b) they are ordered list numbers with the same delimiter (either . or)).

A list is an <u>ordered list</u> if its constituent list items begin with <u>ordered list markers</u>, and a <u>bullet list</u> if its constituent list items begin with <u>bullet list markers</u>.

The <u>start number</u> of an <u>ordered list</u> is determined by the list number of its initial list item. The numbers of subsequent list items are disregarded.

A list is <u>loose</u> if any of its constituent list items are separated by blank lines, or if any of its constituent list items directly contain two block-level elements with a blank line between them. Otherwise a list is <u>tight</u>. (The difference in HTML output is that paragraphs in a loose list are wrapped in $\langle p \rangle$ tags, while paragraphs in a tight list are not.)

Changing the bullet or ordered list delimiter starts a new list:

Example 281

Example 282

In CommonMark, a list can interrupt a paragraph. That is, no blank line is needed to separate a paragraph from a following list:

Example 283

Markdown. p1 does not allow this, through fear of triggering a list via a numeral in a hard-wrapped line:

```
The number of windows in my house is 14. The number of doors is 6.
```

Oddly, though, Markdown. pl does allow a blockquote to interrupt a paragraph, even though the same considerations might apply.

In CommonMark, we do allow lists to interrupt paragraphs, for two reasons. First, it is natural and not uncommon for people to start lists without blank lines:

```
I need to buy
- new shoes
- a coat
- a plane ticket
```

Second, we are attracted to a

principle of uniformity: if a chunk of text has a certain meaning, it will continue to have the same meaning when put into a container block (such as a list item or blockquote).

(Indeed, the spec for $\underline{\text{list items}}$ and $\underline{\text{block quotes}}$ presupposes this principle.) This principle implies that if

```
* I need to buy
- new shoes
- a coat
- a plane ticket
```

is a list item containing a paragraph followed by a nested sublist, as all Markdown implementations agree it is (though the paragraph may be rendered without \(\subseteq \notation \) tags, since the list is "tight"), then

```
I need to buy
- new shoes
- a coat
- a plane ticket
```

by itself should be a paragraph followed by a nested sublist.

Since it is well established Markdown practice to allow lists to interrupt paragraphs inside list items, the <u>principle of uniformity</u> requires us to allow this outside list items as well. (<u>reStructuredText</u> takes a different approach, requiring blank lines before lists even inside other list items.)

In order to solve of unwanted lists in paragraphs with hard-wrapped numerals, we allow only lists starting with ${
m 1}$ to interrupt paragraphs. Thus,

Example 284

```
The number of windows in my house is 14. • The number of doors is 6. The number of windows in my house is 14. • The number of doors is 6.
```

We may still get an unintended result in cases like

Example 285

```
The number of windows in my house is 1. The number of doors is 6. 
 \begin{array}{c} \langle p \rangle \text{The number of windows in my house} \\ \text{is} \langle /p \rangle \\ \langle o1 \rangle \\ \text{<1i>} \text{The number of doors is 6.} \\ \text{</01} \rangle \\ \end{array}
```

but this rule should prevent most spurious list captures.

There can be any number of blank lines between items:

Example 286

Example 287

```
- foo
                                     <u1>
- bar
                                     <1i>foo
••• baz
                                     ul>
                                     <1i>bar
                                     <u1>
····bim
                                     <1i>
                                     \langle p \rangle baz \langle /p \rangle
                                     bim
```

To separate consecutive lists of the same type, or to separate a list from an indented code block that would otherwise be parsed as a subparagraph of the final list item, you can insert a blank HTML comment:

Example 288

Example 289

```
ul>
- •• foo
                                                   <1i>
· · · notcode
                                                   \langle p \rangle foo \langle /p \rangle
                                                   notcode
- •• foo
                                                   <1i>>
<!-->
                                                   \langle p \rangle foo \langle /p \rangle
                                                   ••• code
                                                   <!-- -->
                                                   <code>code
                                                   </code>
```

List items need not be indented to the same level. The following list items will be treated as items at the same list level, since none is indented enough to belong to the previous list item:

Example 290

Example 291

Note, however, that list items may not be indented more than three spaces. Here - e is treated as a paragraph continuation line, because it is indented more than three spaces:

Example 292

And here, 3. c is treated as in indented code block, because it is indented four spaces and preceded by a blank line.

Example 293

This is a loose list, because there is a blank line between two of the list items:

Example 294

```
 \begin{array}{c} - \ a \\ - \ b \\ \end{array} \qquad \begin{array}{c} \langle ul \rangle \\ \langle li \rangle \\ \langle p \rangle a \langle / p \rangle \\ \langle / li \rangle \\ \langle p \rangle b \langle / p \rangle \\ \langle / li \rangle \\ \langle li \rangle \\ \langle p \rangle c \langle / p \rangle \\ \langle / li \rangle \\ \langle | li
```

So is this, with a empty second item:

Example 295

These are loose lists, even though there is no space between the items, because one of the items directly contains two block-level elements with a blank line between them:

Example 296

```
- a
                                                                                       <u1>
- b
                                                                                        <1i>
                                                                                        \langle p \rangle_a \langle /p \rangle
                                                                                       • €
- d
                                                                                       <1i>
                                                                                       \langle p \rangle b \langle /p \rangle
                                                                                        \langle p\rangle_C \langle /p\rangle
                                                                                        <1i>
                                                                                       \langle p \rangle d \langle /p \rangle
```

Example 297

This is a tight list, because the blank lines are in a code block:

Example 298

This is a tight list, because the blank line is between two paragraphs of a sublist. So the sublist is loose while the outer list is tight:

Example 299

This is a tight list, because the blank line is inside the block quote:

Example 300

This list is tight, because the consecutive block elements are not separated by blank lines:

Example 301

A single-paragraph list is tight:

Example 302

Example 303

This list is loose, because of the blank line between the two block elements in the list item:

Example 304

Here the outer list is loose, the inner list tight:

Example 305

Example 306

```
<u1>
• - b
                                                           <1i>>
• = C
                                                           \langle p \rangle_a \langle /p \rangle
                                                           ul>
- d
                                                           \verb|\langle li > b < \! / li >
• = e
                                                           <1i>c</1i>
• - f
                                                           <1i>>
                                                           \langle p\rangle d\langle /p\rangle
                                                           <l
                                                           <1i>e</1i>
                                                           <1i>f</1i>
```

6 Inlines

Inlines are parsed sequentially from the beginning of the character stream to the end (left to right, in left-to-right languages). Thus, for example, in

Example 307

hi is parsed as code, leaving the backtick at the end as a literal backtick.

6.1 Backslash escapes

https://github.github.com/gfm/ 77/138

Any ASCII punctuation character may be backslash-escaped:

Example 308

Backslashes before other characters are treated as literal backslashes:

Example 309

Escaped characters are treated as regular characters and do not have their usual Markdown meanings:

Example 310

```
\*not emphasized*
                                       *not emphasized*
                                       <br/&gt; not a tag
\\br/\> not a tag
\[not a link](/foo)
                                       [not a link](/foo)
\`not code
                                       `not code
1\. not a list
                                       1. not a list
\* not a list
                                       * not a list
\# not a heading
                                       # not a heading
\[foo]: /url "not a reference"
                                       [foo]: /url " not a reference"
\ö not a character entity
                                       ö not a character entity
```

If a backslash is itself escaped, the following character is not:

Example 311

A backslash at the end of the line is a hard line break:

Example 312

foo\ bar	foo bar

Backslash escapes do not work in code blocks, code spans, autolinks, or raw HTML:

Example 313

%[/, .,	<code>\[\`</code>
Example 314	
• • • • • [/]	<pre><code>\[\] </code></pre>

Example 315



Example 316

Example 317

But they work in all other contexts, including URLs and link titles, link references, and <u>info strings</u> in <u>fenced code blocks</u>:

Example 318

Example 319

Example 320

6.2 Entity and numeric character references

Valid HTML entity references and numeric character references can be used in place of the corresponding Unicode character, with the following exceptions:

• Entity and character references are not recognized in code blocks and code spans.

 \mathbf{A}

• Entity and character references cannot stand in place of special characters that define structural elements in CommonMark. For example, although * can be used in place of a literal * character, * cannot replace * in emphasis delimiters, bullet list markers, or thematic breaks.

Conforming CommonMark parsers need not store information about whether a particular character was represented in the source using a Unicode character or an entity reference.

Entity references consist of & + any of the valid HTML5 entity names + ;. The document https://html.spec.whatwg.org/multipage/entities.json is used as an authoritative source for the valid entity references and their corresponding code points.

Example 321

<u>Decimal numeric character references</u> consist of &# + a string of 1–7 arabic digits + ;. A numeric character reference is parsed as the corresponding Unicode character. Invalid Unicode code points will be replaced by the REPLACEMENT CHARACTER (U+FFFD). For security reasons, the code point U+0000 will also be replaced by U+FFFD.

Example 322

```
# Ӓ Ϡ � # Ä Ϡ ♦
```

<u>Hexadecimal numeric character references</u> consist of &# + either X or x + a string of 1-6 hexadecimal digits + x. They too are parsed as the corresponding Unicode character (this time specified with a hexadecimal numeral instead of decimal).

Example 323

Here are some nonentities:

Example 324

Although HTML5 does accept some entity references without a trailing semicolon (such as ©), these are not recognized here, because it makes the grammar too ambiguous:

Example 325

Strings that are not on the list of HTML5 named entities are not recognized as entity references either:

Example 326

```
&MadeUpEntity; &MadeUpEntity;
```

Entity and numeric character references are recognized in any context besides code spans or code blocks, including URLs, <u>link titles</u>, and <u>fenced code block info strings</u>:

Example 327

Example 329

Example 330

Entity and numeric character references are treated as literal text in code spans and code blocks:

Example 331

Example 332

```
<code>f&ouml;f&ouml;
```

Entity and numeric character references cannot be used in place of symbols indicating structure in CommonMark documents.

Example 333

```
      * foo*
      *foo*

      *foo*
      <em>foo</em>

      Example 334
```

Example 335

foo bar	foo
	bar
Example 336	
foo	→foo

Example 337

6.3 Code spans

A <u>backtick string</u> is a string of one or more backtick characters (i) that is neither preceded nor followed by a backtick.

A <u>code span</u> begins with a backtick string and ends with a backtick string of equal length. The contents of the code span are the characters between the two backtick strings, normalized in the following ways:

- First, <u>line endings</u> are converted to <u>spaces</u>.
- If the resulting string both begins and ends with a <u>space</u> character, but does
 not consist entirely of <u>space</u> characters, a single <u>space</u> character is removed
 from the front and back. This allows you to include code that begins or ends
 with backtick characters, which must be separated by whitespace from the
 opening or closing backtick strings.

This is a simple code span:

Example 338

Here two backticks are used, because the code contains a backtick. This example also illustrates stripping of a single leading and trailing space:

Example 339

This example shows the motivation for stripping leading and trailing spaces:

Example 340

```
<code>``</code>
```

Note that only one space is stripped:

Example 341

```
<code> < \( < \) </p>
```

The stripping only happens if the space is on both sides of the string:

Example 342

Only spaces, and not unicode whitespace in general, are stripped in this way:

Example 343

No stripping occurs if the code span contains only spaces:

Example 344

```
<code> </code> </code> </code>
```

Line endings are treated like spaces:

Example 345

```
foo bar • baz</code>
foo bar • baz</code>
```

Example 346

```
foo • <code>foo </code>
```

Interior spaces are not collapsed:

Example 347

```
`foo •• bar • baz</code>
baz`
```

Note that browsers will typically collapse consecutive spaces when rendering <code> elements, so it is recommended that the following CSS be used:

```
code {white-space: pre-wrap;}
```

Note that backslash escapes do not work in code spans. All backslashes are treated literally:

Example 348

Backslash escapes are never needed, because one can always choose a string of n backtick characters as delimiters, where the code does not contain any strings of exactly n backtick characters.

Example 349

foo `` bar `

``foo`bar``	<code>foo`bar</code>
Example 350	

Code span backticks have higher precedence than any other inline constructs except HTML tags and autolinks. Thus, for example, this is not parsed as emphasized text,

<code>foo • ``

bar</code>

Example 351

00`*`	*foo <code>*</code>
-------	---------------------

And this is not parsed as a link:

since the second * is part of a code span:

Example 352

```
[not \ a \ \widehat{\ } link] \ (/foo^{\prime}) \\ \\ \langle p \rangle [not \ a \ \langle code \rangle link] \ (/foo^{\prime}/code^{\prime}) \langle /p \rangle \\
```

Code spans, HTML tags, and autolinks have the same precedence. Thus, this is code:

Example 353

But this is an HTML tag:

Example 354

	<a \"\"\"\"\"\"\"\"\"\"\"\"\"\"\"\"\"\<="" href="\" th=""><th><a "="" href=""></th>	<a "="" href="">
--	---	----------------------

And this is code:

Example 355

But this is an autolink:

Example 356

When a backtick string is not closed by a matching backtick string, we just have literal backticks:

Example 357

The following case also illustrates the need for opening and closing backtick strings to be equal in length:

Example 359

 \mathbf{A}

6.4 Emphasis and strong emphasis

John Gruber's original Markdown syntax description says:

Markdown treats asterisks (*) and underscores (_) as indicators of emphasis. Text wrapped with one * or _ will be wrapped with an HTML \(\subseteq m \) tag; double *'s or _'s will be wrapped with an HTML \(\subseteq strong \) tag.

This is enough for most users, but these rules leave much undecided, especially when it comes to nested emphasis. The original Markdown pl test suite makes it clear that triple *** and ____ delimiters can be used for strong emphasis, and most implementations have also allowed the following patterns:

```
***strong emph***

***strong** in emph*

***emph* in strong**

**in strong *emph***

*in emph **strong***
```

The following patterns are less widely supported, but the intent is clear and they are useful (especially in contexts like bibliography entries):

```
*emph *with emph* in it*
**strong **with strong** in it**
```

Many implementations have also restricted intraword emphasis to the * forms, to avoid unwanted emphasis in words containing internal underscores. (It is best practice to put these in code spans, but users often do not.)

```
internal emphasis: foo*bar*baz
no emphasis: foo bar baz
```

The rules given below capture all of these patterns, while allowing for efficient parsing strategies that do not backtrack.

First, some definitions. A <u>delimiter run</u> is either a sequence of one or more * characters that is not preceded or followed by a non-backslash-escaped * character, or a sequence of one or more characters that is not preceded or followed by a non-backslash-escaped character.

A <u>left-flanking delimiter run</u> is a <u>delimiter run</u> that is (1) not followed by <u>Unicode whitespace</u>, and either (2a) not followed by a <u>punctuation character</u>, or (2b) followed by a <u>punctuation character</u> and preceded by <u>Unicode whitespace</u> or a <u>punctuation character</u>. For purposes of this definition, the beginning and the end of the line count as Unicode whitespace.

A <u>right-flanking delimiter run</u> is a <u>delimiter run</u> that is (1) not preceded by <u>Unicode whitespace</u>, and either (2a) not preceded by a <u>punctuation character</u>, or (2b) preceded by a <u>punctuation character</u> and followed by <u>Unicode whitespace</u> or a <u>punctuation character</u>. For purposes of this definition, the beginning and the end of the line count as Unicode whitespace.

Here are some examples of delimiter runs.

· left-flanking but not right-flanking:

```
***abc
_abc
**"abc"
_"abc"
```

· right-flanking but not left-flanking:

```
abc***
abc_
"abc"**
"abc"_
```

· Both left and right-flanking:

```
abc***def
"abc"_"def"
```

Neither left nor right-flanking:

https://github.github.com/gfm/

```
abc *** def
a _ b
```

(The idea of distinguishing left-flanking and right-flanking delimiter runs based on the character before and the character after comes from Roopesh Chander's <u>vfmd</u>. vfmd uses the terminology "emphasis indicator string" instead of "delimiter run," and its rules for distinguishing left- and right-flanking runs are a bit more complex than the ones given here.)

The following rules define emphasis and strong emphasis:

- 1. A single * character can open emphasis iff (if and only if) it is part of a left-flanking delimiter run.
- 2. A single character <u>can open emphasis</u> iff it is part of a <u>left-flanking delimiter run</u> and either (a) not part of a <u>right-flanking delimiter run</u> or (b) part of a <u>right-flanking delimiter run</u> preceded by punctuation.
- 3. A single * character can close emphasis iff it is part of a right-flanking delimiter run.
- 4. A single character can close emphasis iff it is part of a right-flanking delimiter run and either (a) not part of a left-flanking delimiter run or (b) part of a left-flanking delimiter run followed by punctuation.
- 5. A double ** can open strong emphasis iff it is part of a left-flanking delimiter run.
- 6. A double <u>can open strong emphasis</u> iff it is part of a <u>left-flanking delimiter run</u> and either (a) not part of a <u>right-flanking delimiter run</u> or (b) part of a <u>right-flanking delimiter run</u> preceded by punctuation.
- 7. A double ** can close strong emphasis iff it is part of a right-flanking delimiter run.
- 8. A double <u>can close strong emphasis</u> iff it is part of a <u>right-flanking delimiter run</u> and either (a) not part of a <u>left-flanking delimiter run</u> or (b) part of a <u>left-flanking delimiter run</u> followed by punctuation.
- 9. Emphasis begins with a delimiter that <u>can open emphasis</u> and ends with a delimiter that <u>can close emphasis</u>, and that uses the same character (☐ or ★) as the opening delimiter. The opening and closing delimiters must belong to separate <u>delimiter runs</u>. If one of the delimiters can both open and close emphasis, then the sum of the lengths of the delimiter runs containing the opening and closing delimiters must not be a multiple of 3 unless both lengths are multiples of 3.
- 10. Strong emphasis begins with a delimiter that <u>can open strong emphasis</u> and ends with a delimiter that <u>can close strong emphasis</u>, and that uses the same character (☐ or ※) as the opening delimiter. The opening and closing delimiters must belong to separate <u>delimiter runs</u>. If one of the delimiters can both open and close strong emphasis, then the sum of the lengths of the delimiter runs containing the opening and closing delimiters must not be a multiple of 3 unless both lengths are multiples of 3.

- 11. A literal * character cannot occur at the beginning or end of *-delimited emphasis or **-delimited strong emphasis, unless it is backslash-escaped.
- 12. A literal character cannot occur at the beginning or end of delimited emphasis or delimited strong emphasis, unless it is backslash-escaped.

Where rules 1–12 above are compatible with multiple parsings, the following principles resolve ambiguity:

- 14. An interpretation $\langle em \rangle \langle strong \rangle ... \langle /strong \rangle \langle /em \rangle$ is always preferred to $\langle strong \rangle \langle em \rangle ... \langle /em \rangle \langle /strong \rangle$.
- 15. When two potential emphasis or strong emphasis spans overlap, so that the second begins before the first ends and ends after the first ends, the first takes precedence. Thus, for example, *foo _bar* baz_ is parsed as foo bar baz rather than *foo bar* baz.
- 16. When there are two potential emphasis or strong emphasis spans with the same closing delimiter, the shorter one (the one that opens later) takes precedence. Thus, for example, **foo **bar baz** is parsed as **foo \(\strong \) bar baz \(\strong \) rather than \(\strong \) foo **bar baz \(\strong \).
- 17. Inline code spans, links, images, and HTML tags group more tightly than emphasis. So, when there is a choice between an interpretation that contains one of these elements and one that does not, the former always wins. Thus, for example, *[foo*] (bar) is parsed as *foo* rather than as [foo] (bar).

These rules can be illustrated through a series of examples.

Rule 1:

Example 360

This is not emphasis, because the opening * is followed by whitespace, and hence not part of a Left-flanking delimiter run:

Example 361

This is not emphasis, because the opening * is preceded by an alphanumeric and followed by punctuation, and hence not part of a <u>left-flanking delimiter run</u>:

Example 362

Unicode nonbreaking spaces count as whitespace, too:

Example 363



Intraword emphasis with * is permitted:

Example 364

foo*bar*	foo bar
----------	------------------

Example 365

5*6*78	5 6 78
--------	-----------------

Rule 2:

Example 366

```
_foo bar_ <em>foo bar</em>
```

This is not emphasis, because the opening \blacksquare is followed by whitespace:

Example 367

```
_ foo bar_ _ foo bar_
```

This is not emphasis, because the opening $\ \ \ \$ is preceded by an alphanumeric and followed by punctuation:

Example 368

Example 369

Example 370

5_6_78	5_6_78

Example 371

```
пристаням_стремятся_
я_
```

Here does not generate emphasis, because the first delimiter run is right-flanking and the second left-flanking:

Example 372

This is emphasis, even though the opening delimiter is both left- and right-flanking, because it is preceded by punctuation:

Example 373

Rule 3:

This is not emphasis, because the closing delimiter does not match the opening delimiter:

Example 374



This is not emphasis, because the closing * is preceded by whitespace:

Example 375

A newline also counts as whitespace:

Example 376



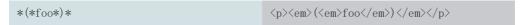
This is not emphasis, because the second * is preceded by punctuation and followed by an alphanumeric (hence it is not part of a <u>right-flanking delimiter run</u>:

Example 377



The point of this restriction is more easily appreciated with this example:

Example 378



Intraword emphasis with * is allowed:

Example 379

*foo*bar	foo bar
----------	------------------

Rule 4:

This is not emphasis, because the closing is preceded by whitespace:

Example 380

This is not emphasis, because the second is preceded by punctuation and followed by an alphanumeric:

Example 381



This is emphasis within emphasis:

Example 382



Intraword emphasis is disallowed for _:

Example 383

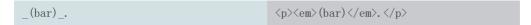
Example 384

Example 385



This is emphasis, even though the closing delimiter is both left- and right-flanking, because it is followed by punctuation:

Example 386



Rule 5:

Example 387

foo bar	foo bar
-------------	--------------------------

This is not strong emphasis, because the opening delimiter is followed by whitespace:

Example 388

This is not strong emphasis, because the opening ** is preceded by an alphanumeric and followed by punctuation, and hence not part of a <u>left-flanking delimiter run</u>:

Example 389

Intraword strong emphasis with ** is permitted:

Example 390

foo**bar**	foo**bar**
------------	------------

Rule 6:

Example 391

```
__foo bar__ <strong>foo bar</strong>
```

This is not strong emphasis, because the opening delimiter is followed by whitespace:

Example 392

```
__ foo bar__ __ foo bar__
```

A newline counts as whitespace:

Example 393

This is not strong emphasis, because the opening __ is preceded by an alphanumeric and followed by punctuation:

Example 394

Intraword strong emphasis is forbidden with ___:

Example 395

```
foo_bar__ foo_bar__
```

Example 396

Example 397

```
пристаням__стремятся_     пристаням__стремятс
____
```

Example 398

This is strong emphasis, even though the opening delimiter is both left- and right-flanking, because it is preceded by punctuation:

Example 399

Rule 7:

This is not strong emphasis, because the closing delimiter is preceded by whitespace:

Example 400

(Nor can it be interpreted as an emphasized *foo bar *, because of Rule 11.)

This is not strong emphasis, because the second ** is preceded by punctuation and followed by an alphanumeric:

Example 401



The point of this restriction is more easily appreciated with these examples:

Example 402



Example 403

```
**Gomphocarpus (*Gomphocarpus physocarpus (<em>Gomphocarpus physocarpus physicarpus physic
```

Example 404

Intraword emphasis:

Example 405

foobar	foo bar
------------	--------------------------

Rule 8:

This is not strong emphasis, because the closing delimiter is preceded by whitespace:

Example 406



This is not strong emphasis, because the second __ is preceded by punctuation and followed by an alphanumeric:

Example 407



The point of this restriction is more easily appreciated with this example:

Example 408

Intraword strong emphasis is forbidden with ::

Example 409

```
__foo__bar
```

Example 410

```
__пристаням__стремятс      __пристаням__стремят      ся
```

Example 411

This is strong emphasis, even though the closing delimiter is both left- and right-flanking, because it is followed by punctuation:

Example 412

(bar)	(bar) .
-------	--------------------------

Rule 9:

Any nonempty sequence of inline elements can be the contents of an emphasized span.

Example 413

Example 414

*foo	foo
bar*	bar

In particular, emphasis and strong emphasis can be nested inside emphasis:

Example 415

_foo •_bar baz_	foo bar •
	baz
	oda (, cm/ (, p/

Example 416

foo •bar baz_	foo bar baz
	-

Example 417

Example 418

*foo *bar**	foo bar
-------------	---------------------------

Example 419

*foo **bar** baz*	foo bar •
	baz
	baz v, cm/ v, p/

Example 420

*foo**bar**baz*	foobarbaz
	\/ p/

Note that in the preceding case, the interpretation

```
<em>foo</em><em>bar<em></em>baz</em>
```

is precluded by the condition that a delimiter that can both open and close (like the * after foo) cannot form emphasis if the sum of the lengths of the delimiter runs containing the opening and closing delimiters is a multiple of 3 unless both lengths are multiples of 3.

For the same reason, we don't get two consecutive emphasis sections in this example:

Example 421

*foo**bar*	foo**bar
------------	-------------------

The same condition ensures that the following cases are all strong emphasis nested inside emphasis, even when the interior spaces are omitted:

Example 422

***foo** bar*	foo bar
	7.1
Example 423	
*foo **bar***	foo bar
	// P/
Example 424	
*foo**bar***	foobar

When the lengths of the interior closing and opening delimiter runs are *both* multiples of 3, though, they can match to create emphasis:

Example 425

foo***bar***baz	foo bar baz
Example 426	
foo*****bar*****baz	foo
	<pre>***baz</pre>

Indefinite levels of nesting are possible:

Example 427

*foo **bar *baz* bim** bop*	foo bar baz = bim bop
Example 428	,

*foo [*bar*](/url)*	foo bar
	(, a) (, cm) (, p)

There can be no empty emphasis or strong emphasis:

Example 429

** is not an empty emphasis	*** is not an empty emphasis
Example 430	
**** is not an empty strong emphasis	**** is not an empty strong • emphasis
	embirasis// h/

Rule 10:

Any nonempty sequence of inline elements can be the contents of an strongly emphasized span.

Example 431

foo [bar](/url)	foo bar
	\/ Stiong/\/ p/

Example 432

**foo	foo
bar**	bar

In particular, emphasis and strong emphasis can be nested inside strong emphasis:

Example 433

foo _bar_ baz	foo bar • baz
	baz (belong) (p)

Example 434

foo •bar baz	foo bar = baz
	baz // strong/ // p/

Example 435

foo bar	foo • bar
	bar y berong, y p

Example 436

foo **bar*	foo bar
Example 437	
foo *bar* baz	foo bar • baz
Example 438	
foo*bar*baz	foobarbaz
Example 439	
***foo* bar**	foo bar
Example 440	
foo *bar*	foo bar

Indefinite levels of nesting are possible:

Example 441

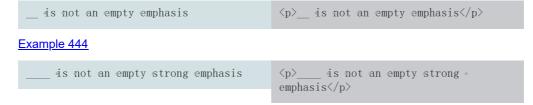
foo *bar **baz bim* bop**	foo bar baz bim bop
Example 442	

**foo [*bar*] (/url) **

foo
bar

There can be no empty emphasis or strong emphasis:

Example 443



Rule 11:

Example 445

Exam	n	ما	44	16
LAGIII	ש	U		TU

foo ***	foo *
Example 447	
foo *_*	$\langle p \rangle$ foo $\langle em \rangle_{\langle em \rangle \langle p \rangle}$
Example 448	
foo ****	foo *****
Example 449	
foo *****	foo *
Example 450	
foo **_**	foo _

Note that when delimiters do not match evenly, Rule 11 determines that the excess literal * characters will appear outside of the emphasis, rather than inside it:

Example 451

<u>Example 401</u>		
**foo*	* foo	
Example 452		
*foo**	foo *	
Example 453		
***f00**	* foo	
Example 454		
****f00*	*** foo	
Example 455		
foo*	foo *	
Example 456		
*foo****	foo ***	

Rule 12:

Example 457

Example 458

foo <u>_</u> \	$\langle p \rangle$ foo $\langle em \rangle_{\langle em \rangle \langle p \rangle}$
Example 459	
foo <u>*</u> *_	foo *
Example 460	
foo •	⟨p⟩foo •⟨/p⟩
Example 461	
foo •\	$\label{eq:pfoodstrong} $$\langle p\rangle$ foo $$\langle strong \rangle_{\ \ \ }$$
Example 462	
foo •_*_	foo *
Example 463	
foo_	_ foo

Note that when delimiters do not match evenly, Rule 12 determines that the excess literal _ characters will appear outside of the emphasis, rather than inside it:

Example 464

_foo	foo _
Example 465	
foo	$\label{eq:cong} $$ \p>_{strong}\foo$
Example 466	
foo_	$\langle p \rangle_{\underline{\hspace{1cm}}} \langle em \rangle foo \langle /em \rangle \langle /p \rangle$
Example 467	
foo	foo _
Example 468	
_foo	foo

Rule 13 implies that if you want emphasis nested directly inside emphasis, you must use different delimiters:

Example 469

foo	$\ensuremath{\scriptsize \langle p\rangle\langle strong\rangle foo\langle/strong\rangle\langle/p\rangle}$
---------	---

Example 470

foo			
Example 471			
foo	foo		
Example 472			
*foo*	foo		

However, strong emphasis within strong emphasis is possible without switching delimiters:

Example 473

f00	foo
	(/ 3 t1 0 lig/ \/ p/

Example 474

foo	foo
	// Strong/// p/

Rule 13 can be applied to arbitrarily long sequences of delimiters:

Example 475

*****f00*****	foo
	Vacious, vacious, vab.

Rule 14:

Example 476

f00	foo

Example 477

foo	foo
	(/ B C C O N B / C M / P /

Rule 15:

Example 478

	
foo _bar baz_	foo _bar baz_
Example 479	
*foobar *baz bim bam*	foo bar *baz • bim bam

Rule 16:

Example 480

foo **bar baz	**foo bar baz

Example 481

*foo *bar baz* *f	oo bar baz
-------------------	---------------------

Rule 17:

Example 482

[bar](/url)	* bar*
---------------	---------------------------

Example 483

foo [bar] (/url)	`>bar_
--------------------	--------

Example 484

	* 	%img src="foo" title="*"/>
--	------------------------------	----------------------------

Example 485

** 	**

Example 486

Example 487

*a • * *	a <code>*</code>
----------	---------------------------

Example 488

_a • _ `_	a <code>_</code>
-----------	---------------------------

Example 489

a <http: ?q="" foo.bar=""></http:>	**a href="http://foo.bar/?
	q=**">http://foo.bar/?q=**
	q / 11 ctp. / / 100; bal / . q / p/

Example 490

a <http: ?q="" foo.bar=""></http:>	_a <a href="http://foo.bar/?
q=">http://foo.bar/?q=
	q /11ttp://100.bai/.q \/ a/ \/ p/

6.5 Strikethrough (extension)

GFM enables the strikethrough extension, where an additional emphasis type is available.

Strikethrough text is any text wrapped in two tildes (~).

Example 491

```
~~Hi~~ Hello, world! <del>Hi</del> Hello, world!
```

As with regular emphasis delimiters, a new paragraph will cause strikethrough parsing to cease:

Example 492

6.6 Links

A link contains <u>link text</u> (the visible text), a <u>link destination</u> (the URI that is the link destination), and optionally a <u>link title</u>. There are two basic kinds of links in Markdown. In <u>inline links</u> the destination and title are given immediately after the link text. In <u>reference links</u> the destination and title are defined elsewhere in the document.

A <u>link text</u> consists of a sequence of zero or more inline elements enclosed by square brackets ([and]). The following rules apply:

- Links may not contain other links, at any level of nesting. If multiple otherwise valid link definitions appear nested inside each other, the inner-most definition is used.
- Brackets are allowed in the <u>link text</u> only if (a) they are backslash-escaped or
 (b) they appear as a matched pair of brackets, with an open bracket [, a
 sequence of zero or more inlines, and a close bracket].
- Backtick <u>code spans</u>, <u>autolinks</u>, and raw <u>HTML tags</u> bind more tightly than the brackets in link text. Thus, for example, [foo`]` could not be a link text, since the second I is part of a code span.
- The brackets in link text bind more tightly than markers for emphasis and strong emphasis. Thus, for example, *[foo*] (ur1) is a link.

A link destination consists of either

- a sequence of zero or more characters between an opening < and a closing > that contains no line breaks or unescaped < or > characters, or
- a nonempty sequence of characters that does not start with \(\), does not include ASCII space or control characters, and includes parentheses only if (a) they are backslash-escaped or (b) they are part of a balanced pair of unescaped parentheses. (Implementations may impose limits on parentheses nesting to avoid performance issues, but at least three levels of nesting should be supported.)

A link title consists of either

- a sequence of zero or more characters between straight double-quote characters ("), including a " character only if it is backslash-escaped, or
- a sequence of zero or more characters between straight single-quote characters ('), including a character only if it is backslash-escaped, or
- a sequence of zero or more characters between matching parentheses ((...)), including a (or) character only if it is backslash-escaped.

Although link titles may span multiple lines, they may not contain a blank line.

An <u>inline link</u> consists of a <u>link text</u> followed immediately by a left parenthesis (, optional <u>whitespace</u>, an optional <u>link destination</u>, an optional <u>link title</u> separated from the link destination by <u>whitespace</u>, optional <u>whitespace</u>, and a right parenthesis). The link's text consists of the inlines contained in the <u>link text</u> (excluding the enclosing square brackets). The link's URI consists of the link destination, excluding enclosing <... if present, with backslash-escapes in effect as described above. The link's title consists of the link title, excluding its enclosing delimiters, with backslash-escapes in effect as described above.

Here is a simple inline link:

Example 493

[link](/uri "title")	<a href="/uri" •<br="">title="title">link
	title title / link \/ d/ \/ p/

The title may be omitted:

Example 494

[link](/uri)	link
--------------	-------------------------

Both the title and the destination may be omitted:

Example 495

[link]()	link
Example 496	
[link](<>)	link

The destination can only contain spaces if it is enclosed in pointy brackets:

Example 497

[link](/my uri)	[link] (/my uri)
Example 498	
[link]()	link

The destination cannot contain line breaks, even if enclosed in pointy brackets:

Example 499

[link](foo bar)	[link] (foo bar)

Example 500

[link](<foo bar>)</foo 	[link] (<foo bar>)</foo

The destination can contain) if it is enclosed in pointy brackets:

Example 501

```
[a](\( \b) c \( \rangle \) c"\( \rangle a \) \( \rangle a \) \
```

Pointy brackets that enclose links must be unescaped:

Example 502

These are not links, because the opening pointy bracket is not matched properly:

Example 503

[a] (<b) c<br="">[a] (<b) c=""> [a] (c)</b)></b)>	[a] (<b) (<b)="" [a]="" c=""> [a] (c)</b)>
[4] (10/0)	[4] (16/6) (/ p/

Parentheses inside the link destination may be escaped:

Example 504

Any number of parentheses are allowed without escaping, as long as they are balanced:

Example 505

However, if you have unbalanced parentheses, you need to escape or use the $\langle \ldots \rangle$ form:

105/138

Example 506

https://github.github.com/gfm/

[link](foo\(and\(bar\))	link
Example 507	
[link](<foo(and(bar)>)</foo(and(bar)>	link

Parentheses and other symbols can also be escaped, as usual in Markdown:

Example 508

A link can contain fragment identifiers and gueries:

Example 509

Note that a backslash before a non-escapable character is just a backslash:

Example 510

URL-escaping should be left alone inside the destination, as all URL-escaped characters are also valid URL characters. Entity and numerical character references in the destination will be parsed into the corresponding Unicode code points, as usual. These may be optionally URL-escaped when written as HTML, but this spec does not enforce any particular policy for rendering URLs in HTML or other formats. Renderers may make different decisions about how to escape or normalize URLs in the output.

Example 511

Note that, because titles can often be parsed as destinations, if you try to omit the destination and keep the title, you'll get unexpected results:

Example 512

Titles may be in single quotes, double quotes, or parentheses:

Example 513

```
[link](/url "title")

[link](/url 'title')

[link](/url '(title))

(a href="/url" title="title">link</a>
(a href="/url" title="title">link</a>
(b)
```

Backslash escapes and entity and numeric character references may be used in titles:

Example 514

Titles must be separated from the link using a <u>whitespace</u>. Other <u>Unicode whitespace</u> like non-breaking space doesn't work.

Example 515

Nested balanced quotes are not allowed without escaping:

Example 516

But it is easy to work around this by using a different quote type:

Example 517

(Note: Markdown. p1 did allow double quotes inside a double-quoted title, and its test suite included a test demonstrating this. But it is hard to see a good rationale for the extra complexity this brings, since there are already many ways—backslash escaping, entity and numeric character references, or using a different quote type for the enclosing title—to write titles containing double quotes. Markdown. p1's handling of titles has a number of other strange features. For example, it allows single-quoted titles in inline links, but not reference links. And, in reference links but not inline links, it allows a title to begin with and end with . Markdown. p1 1.0.1 even allows titles with no closing quotation mark, though 1.0.2b8 does not. It seems preferable to adopt a simple, rational rule that works the same way in inline links and link reference definitions.)

Whitespace is allowed around the destination and title:

Example 518

[link](••/uri	<a href="/uri" th="" ⋅<="">
•"title" •)	title="title">link

But it is not allowed between the link text and the following parenthesis:

Example 519

The link text may contain balanced brackets, but not unbalanced ones, unless they are escaped:

Example 520

[link [foo [bar]]](/uri)	link [foo [bar]]
	⟨/p⟩
	√ P′

Example 521

ink] bar](/uri)	[link] bar](/uri)
-----------------	-------------------

Example 522

Example 523

[link \[bar] (/uri)

The link text may contain inline content:

Example 524

[link *foo **bar** `#`*](/uri)	link foo

Example 525

<img p="" src="moon.jpg" •<=""/>
alt="moon" />
are moon , , , ar , p

However, links may not contain other links, at any level of nesting.

Example 526



Example 527

[foo *[bar [baz](/uri)](/uri)*](/uri)	[foo [bar baz] (/uri)] (/uri)
	<pre></pre>

These cases illustrate the precedence of link text grouping over emphasis grouping:

Example 529

[foo](/uri)	* foo*
Example 530	
[foo *bar](baz*)	foo *bar

Note that brackets that aren't part of links do not take precedence:

Example 531

These cases illustrate the precedence of HTML tags, code spans, and autolinks over link grouping:

Example 532

[foo <bar attr="](baz)"></bar>	[foo ⟨bar attr="](baz)">⟨/p⟩	
Example 533		
[foo`](/uri)`	[foo <code>] (/uri)</code>	
Example 534		
[foo <http: ?search="](uri)" example.com=""></http:>	\[foo\a href="http://example.com/? search=\%5D(uri)"\rangle.com/? search=\] (uri)\(\langle \rangle \rangle \rangle)\(\rangle \rangle \rangle \rangle \rangle \rangle.\)	

There are three kinds of reference links: full, collapsed, and shortcut.

A <u>full reference link</u> consists of a <u>link text</u> immediately followed by a <u>link label</u> that <u>matches</u> a <u>link reference definition</u> elsewhere in the document.

A <u>link label</u> begins with a left bracket ([) and ends with the first right bracket ([) that is not backslash-escaped. Between these brackets there must be at least one <u>non-whitespace character</u>. Unescaped square bracket characters are not allowed inside the opening and closing square brackets of <u>link labels</u>. A link label can have at most 999 characters inside the square brackets.

One label <u>matches</u> another just in case their normalized forms are equal. To normalize a label, strip off the opening and closing brackets, perform the *Unicode case fold*, strip leading and trailing <u>whitespace</u> and collapse consecutive internal <u>whitespace</u> to a single space. If there are multiple matching reference link definitions, the one that comes first in the document is used. (It is desirable in such cases to emit a warning.)

The link's URI and title are provided by the matching link reference definition.

Here is a simple example:

Example 535

The rules for the <u>link text</u> are the same as with <u>inline links</u>. Thus:

The link text may contain balanced brackets, but not unbalanced ones, unless they are escaped:

Example 536

Example 537

The link text may contain inline content:

Example 538

Example 539

However, links may not contain other links, at any level of nesting.

Example 540

[foo [bar](/uri)][ref]	[foo bar] <a th="" •<="">
[ref]: √uri	href="/uri">ref
[ICI]. / uII	

(In the examples above, we have two <u>shortcut reference links</u> instead of one $\underline{\text{full}}$ reference link.)

The following cases illustrate the precedence of link text grouping over emphasis grouping:

Example 542

Example 543

These cases illustrate the precedence of HTML tags, code spans, and autolinks over link grouping:

Example 544

```
[foo <bar attr="][ref]"> [foo <bar attr="][ref]">
[ref]: /uri
```

Example 545

[foo`][ref]`	$\label{eq:pocode} $$ \p>[foo][ref]$
[ref]: √uri	

Example 546



Matching is case-insensitive:

Example 547

Unicode case fold is used:

Example 548

```
[ß] <a href="/url">β</a>
[SS]: √url
```

Consecutive internal <u>whitespace</u> is treated as one space for purposes of determining matching:

Example 549

No whitespace is allowed between the link text and the link label:

Example 550

Example 551

This is a departure from John Gruber's original Markdown syntax description, which explicitly allows whitespace between the link text and the link label. It brings reference links in line with inline links, which (according to both original Markdown and this spec) cannot have whitespace after the link text. More importantly, it prevents inadvertent capture of consecutive shortcut reference links. If whitespace is allowed between the link text and the link label, then in the following we will have a single reference link, not two shortcut reference links, as intended:

[foo] [bar] [foo]: /url1 [bar]: /url2

https://github.github.com/gfm/

(Note that <u>shortcut reference links</u> were introduced by Gruber himself in a beta version of <u>Markdown. pl</u>, but never included in the official syntax description. Without shortcut reference links, it is harmless to allow space between the link text and link label; but once shortcut references are introduced, it is too dangerous to allow this, as it frequently leads to unintended results.)

When there are multiple matching <u>link reference definitions</u>, the first is used:

Example 552

Note that matching is performed on normalized strings, not parsed inline content. So the following does not match, even though the labels define equivalent inline content:

Example 553

Link labels cannot contain brackets, unless they are backslash-escaped:

Example 554

Example 555

Example 556

Example 557

Note that in this example] is not backslash-escaped:

Example 558

A link label must contain at least one non-whitespace character:

Example 559

Example 560

A <u>collapsed reference link</u> consists of a <u>link label</u> that <u>matches</u> a <u>link reference definition</u> elsewhere in the document, followed by the string []. The contents of the first link label are parsed as inlines, which are used as the link's text. The link's URI and title are provided by the matching reference link definition. Thus, [foo][] is equivalent to [foo][foo].

Example 561

Example 562

The link labels are case-insensitive:

Example 563

As with full reference links, whitespace is not allowed between the two sets of brackets:

Example 564

A **shortcut reference link** consists of a <u>link label</u> that <u>matches</u> a <u>link reference definition</u> elsewhere in the document and is not followed by [] or a link label. The contents of the first link label are parsed as inlines, which are used as the link's text. The link's URI and title are provided by the matching link reference definition. Thus, [foo] is equivalent to [foo][].

Example 565

Example 566

Example 567

Example 568

The link labels are case-insensitive:

Example 569

A space after the link text should be preserved:

Example 570

If you just want bracketed text, you can backslash-escape the opening bracket to avoid links:

Example 571

Note that this is a link, because a link label ends with the first following closing bracket:

Example 572

Full and compact references take precedence over shortcut references:

Example 573

Example 574

Inline links also take precedence:

Example 575

Example 576

In the following case [bar][baz] is parsed as a reference, [foo] as normal text:

Example 577

```
[foo][bar][baz]
                                       [foo] <a href="/url">bar</a>
[baz]: /url
```

Here, though, [foo] [bar] is parsed as a reference, since [bar] is defined:

Example 578

```
[foo][bar][baz]
                                        <a href="/url2">foo</a><a
                                       href="/url1">baz</a>
[baz]: √url1
[bar]: /url2
```

Here [foo] is not parsed as a shortcut reference, because it is followed by a link label (even though [bar] is not defined):

Example 579

```
\langle p \rangle [foo] \langle a href="/url1" \rangle bar \langle /a \rangle \langle /p \rangle
[foo][bar][baz]
[baz]: /url1
[foo]: /url2
```

6.7 Images

Syntax for images is like the syntax for links, with one difference. Instead of link text, we have an image description. The rules for this are the same as for link text, except that (a) an image description starts with ! rather than , and (b) an image description may contain links. An image description has inline elements as its contents. When an image is rendered to HTML, this is standardly used as the image's alt attribute.

Example 580

```
<img src="/url" alt="foo" title="title" />
 ![foo](/url "title")
Example 581
```

```
<img src="train.jpg" alt="foo bar" |
![foo *bar*]
                                       title="train & tracks" />
[foo *bar*]: train.jpg "train & tracks"
```

Example 582

```
![foo ![bar](/url)](/url2)
                                      <img src="/url2" alt="foo bar" />
```

Example 583

```
![foo [bar](/url)](/url2)
                                                  <img src="/url2" alt="foo bar" />
                                                  \langle /p \rangle
```

https://github.github.com/gfm/ 117/138 Though this spec is concerned with parsing, not rendering, it is recommended that in rendering to HTML, only the plain string content of the <u>image description</u> be used. Note that in the above example, the alt attribute's value is foo bar, not foo [bar] (/ur1) or foo (a href="/ur1">bar. Only the plain string content is rendered, without formatting.

Example 584

Example 585

Example 586

![foo](train.jpg)	
	⟨/p⟩
	\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \

Example 587

My ![foo bar](/path/to/train.jpg	My <img src="/path/to/train.jpg" th="" •<=""/>
"title" •••)	alt="foo bar" title="title" />

Example 588

![foo](<url>)</url>	

Example 589

	p>
-----------	----

Reference-style:

Example 590

![foo][bar]	
[bar]: /url	

Example 591

![foo][bar]	
[BAR]: /url	

Collapsed:

Example 592

The labels are case-insensitive:

Example 594

As with reference links, whitespace is not allowed between the two sets of brackets:

Example 595

Shortcut:

Example 596

Example 597

Note that link labels cannot contain unescaped brackets:

Example 598

The link labels are case-insensitive:

Example 599

If you just want a literal I followed by bracketed text, you can backslash-escape the opening I:

Example 600

If you want a link after a literal !, backslash-escape the !:

Example 601

6.8 Autolinks

<u>Autolink</u>s are absolute URIs and email addresses inside

and

. They are parsed as links, with the URL or email address as the link label.

A <u>URI autolink</u> consists of <, followed by an <u>absolute URI</u> followed by >. It is parsed as a link to the URI, with the URI as the link's label.

An <u>absolute URI</u>, for these purposes, consists of a <u>scheme</u> followed by a colon (:) followed by zero or more characters other than ASCII <u>whitespace</u> and control characters, \le , and \ge . If the URI includes these characters, they must be percentencoded (e.g. %20 for a space).

For purposes of this spec, a **scheme** is any sequence of 2–32 characters beginning with an ASCII letter and followed by any combination of ASCII letters, digits, or the symbols plus ("+"), period ("."), or hyphen ("-").

Here are some valid autolinks:

Example 602

<http: foo.bar.baz=""></http:>	http://foo.ba
	r. baz

Example 603

<pre><irc: baz="" foo.bar:2233=""></irc:></pre>	<a *="" href="irc://foo.bar:2233/baz">irc://foo
	. bar:2233/baz

Uppercase is also fine:

Example 605

<mailto:f00@bar.baz></mailto:f00@bar.baz>	MATLTO:F00@BA
	R. BAZ

Note that many strings that count as <u>absolute URIs</u> for purposes of this spec are not valid URIs, because their schemes are not registered or because of other problems with their syntax:

Example 606

<a+b+c:d></a+b+c:d>	p>
---------------------	----

Example 607

<made-up-scheme: foo,bar=""></made-up-scheme:>	<a href="made-up-
scheme://foo,bar">made-up-
	scheme://foo,bar

Example 608

<http:></http:>	http:///

Example 609

<localhost:5001 foo=""></localhost:5001>	<a =="" href="localhost:5001/foo">localhost:500
	1/foo

Spaces are not allowed in autolinks:

Example 610

http://foo.bar/baz bim	%lt;http://foo.bar/baz bim>
---	-----------------------------

Backslash-escapes do not work inside autolinks:

Example 611

lem:lem:lem:lem:lem:lem:lem:lem:lem:lem:	htt
	p://example.com/\[\

An <u>email autolink</u> consists of \le , followed by an <u>email address</u>, followed by \ge . The link's label is the email address, and the URL is <u>mailto</u>: followed by the email address.

An $\underline{\text{email address}}$, for these purposes, is anything that matches the $\underline{\text{non-normative}}$ $\underline{\text{regex from the HTML5 spec}}$:

Examples of email autolinks:

Example 612

<foo@bar.example.com></foo@bar.example.com>	foo@b
	ar.example.com

Example 613

<pre><foo+special@bar.baz-bar0.com></foo+special@bar.baz-bar0.com></pre>	foo+special@Bar.baz-bar0.com">foo+special@Bar.baz-bar0.com">foo+special@Bar.baz-bar0.com">foo+special@Bar.baz-bar0.com">foo+special@Bar.baz-bar0.com">foo+special@Bar.baz-bar0.com
	bar0. com

Backslash-escapes do not work inside email autolinks:

Example 614

<foo\+@bar.example.com></foo\+@bar.example.com>	$\protect\$ foo+@bar.example.com> $\protect\$
---	---

These are not autolinks:

Example 615

\Diamond	&1t;>
Example 616	
< http://foo.bar →	%lt; http://foo.bar >
Example 617	
<m:abc></m:abc>	%1t;m:abc>
Example 618	
<foo.bar.baz></foo.bar.baz>	%lt;foo.bar.baz>
Example 619	
http://example.com	http://example.com

Example 620

foo@bar.example.com foo@bar.example.com

6.9 Autolinks (extension)

GFM enables the autolink extension, where autolinks will be recognised in a greater number of conditions.

<u>Autolinks</u> can also be constructed without requiring the use of \langle and to \rangle to delimit them, although they will be recognized under a smaller set of circumstances. All such recognized autolinks can only come at the beginning of a line, after whitespace, or any of the delimiting characters *, $_{\sim}$, $_{\sim}$, and $_{\sim}$.

An **extended www autolink** will be recognized when the text www. is found followed by a <u>valid domain</u>. A <u>valid domain</u> consists of segments of alphanumeric characters, underscores (_) and hyphens (-) separated by periods (.). There must be at least one period, and no underscores may be present in the last two segments of the domain.

The scheme http will be inserted automatically:

Example 621

www.commonmark.org	www.co
	mmonmark.org

After a valid domain, zero or more non-space non
√ characters may follow:

Example 622

We then apply extended autolink path validation as follows:

Trailing punctuation (specifically, ?, !, ., ,, ., *, ., and $\tilde{}$) will not be considered part of the autolink, though they may be included in the interior of the link:

Example 623

When an autolink ends in), we scan the entire autolink for the total number of parentheses. If there is a greater number of closing parentheses than opening ones,

we don't consider the unmatched trailing parentheses part of the autolink, in order to facilitate including an autolink inside a parenthesis:

Example 624

```
www.google.com/search?q=Markup+
                                         <a •
                                         href="http://www.google.com/search?
(business)
                                         q=Markup+
www.google.com/search?q=Markup+
                                         (business)">www.google.com/search?
(business)))
                                         q=Markup+(business)</a>
                                         <a
                                         href="http://www.google.com/search?
(www.google.com/search?q=Markup+
(business))
                                         q=Markup+
                                         (business)">www.google.com/search?
(www.google.com/search?q=Markup+
                                         q=Markup+(business)</a>))
(business)
                                         (<a
                                         href="http://www.google.com/search?
                                         q=Markup+
                                         (business)">www.google.com/search?
                                         q=Markup+(business)</a>)
                                         (<a
                                         href="http://www.google.com/search?
                                         q=Markup+
                                         (business)">www.google.com/search?
                                         q=Markup+(business)</a>
```

This check is only done when the link ends in a closing parentheses), so if the only parentheses are in the interior of the autolink, no special rules are applied:

Example 625

If an autolink ends in a semicolon (;), we check to see if it appears to resemble an entity reference; if the preceding text is & followed by one or more alphanumeric characters. If so, it is excluded from the autolink:

Example 626

immediately ends an autolink.

Example 627

An <u>extended url autolink</u> will be recognised when one of the schemes http://, or https://, followed by a <u>valid domain</u>, then zero or more non-space non-< characters according to extended autolink path validation:

Example 628

An <u>extended email autolink</u> will be recognised when an email address is recognised within any text node. Email addresses are recognised according to the following rules:

- One ore more characters which are alphanumeric, or ., -, _, or +.
- An @ symbol.
- One or more characters which are alphanumeric, or or _, separated by periods (.). There must be at least one period. The last character must not be one of or _.

The scheme mailto: will automatically be added to the generated link:

Example 629

foo@bar.baz	foo@bar.baz<!--</th-->
	a>

+ can occur before the @, but not after.

Example 630

., -, and _ can occur on both sides of the @, but only . may occur at the end of the email address, in which case it will not be considered part of the address:

Example 631

6.10 Raw HTML

Text between

and

that looks like an HTML tag is parsed as a raw HTML tag and will be rendered in HTML without escaping. Tag and attribute names are not limited to current HTML tags, so custom tags (and even, say, DocBook tags) may be used.

Here is the grammar for tags:

A <u>tag name</u> consists of an ASCII letter followed by zero or more ASCII letters, digits, or hyphens (-).

An <u>attribute</u> consists of <u>whitespace</u>, an <u>attribute name</u>, and an optional <u>attribute value specification</u>.

An <u>attribute name</u> consists of an ASCII letter, _, or :, followed by zero or more ASCII letters, digits, _, ., :, or -. (Note: This is the XML specification restricted to ASCII. HTML5 is laxer.)

An <u>attribute value specification</u> consists of optional <u>whitespace</u>, a character, optional <u>whitespace</u>, and an <u>attribute value</u>.

An <u>attribute value</u> consists of an <u>unquoted attribute value</u>, a <u>single-quoted attribute value</u>, or a <u>double-quoted attribute value</u>.

A <u>single-quoted attribute value</u> consists of ', zero or more characters not including ', and a final '.

A <u>double-quoted attribute value</u> consists of $\tilde{\ }$, zero or more characters not including $\tilde{\ }$, and a final $\tilde{\ }$.

A <u>closing tag</u> consists of the string $\langle \cdot \rangle$, a <u>tag name</u>, optional <u>whitespace</u>, and the character \rangle .

An <u>HTML comment</u> consists of $\langle !-- + text + -- \rangle$, where *text* does not start with \rangle or $- \rangle$, does not end with $- \rangle$, and does not contain $- - \rangle$. (See the <u>HTML5 spec.</u>)

A <u>processing instruction</u> consists of the string $\langle ? \rangle$, a string of characters not including the string $? \rangle$, and the string $? \rangle$.

A <u>declaration</u> consists of the string $\langle \cdot \cdot \rangle$, a name consisting of one or more uppercase ASCII letters, <u>whitespace</u>, a string of characters not including the character \geq , and the character \geq .

A <u>CDATA section</u> consists of the string $\langle ! [CDATA[], a string of characters not including the string]]>, and the string]]>.$

An <u>HTML tag</u> consists of an <u>open tag</u>, a <u>closing tag</u>, an <u>HTML comment</u>, a <u>processing instruction</u>, a <u>declaration</u>, or a <u>CDATA section</u>.

Here are some simple open tags:

Example 632

<a><bab><c2c></c2c></bab>	<a><bab><c2c></c2c></bab>
-------------------------------	-------------------------------

Empty elements:

Example 633

⟨a/⟩⟨b2/⟩	<a> <b2></b2>

Whitespace is allowed:

Example 634

With attributes:

Example 635

Custom tag names can be used:

Example 636

Illegal tag names, not parsed as HTML:

Example 637

Illegal attribute names:

Example 638

<a h*#ref="hi"> <a h*#ref=&quo	t;hi">
---	--------

Illegal attribute values:

Example 639

Illegal whitespace:

Example 640

Missing whitespace:

Example 641

Closing tags:

Example 642

<	/foo
---	------

Illegal attributes in closing tag:

Example 643

Comments:

Example 644

Example 645

```
foo <!-- not a comment -- two hyphens -- <p>foo &lt;!-- not a comment -- two hyphens -- &gt;
```

Not comments:

Example 646

Processing instructions:

Example 647

foo php echo \$a; ?	foo php echo \$a; ?
---------------------	---------------------

Declarations:

Example 648

CDATA sections:

Example 649

Entity and numeric character references are preserved in HTML attributes:

Example 650

```
foo <a href="%ouml;"> foo <a href="%ouml;">
```

Backslash escapes do not work in HTML attributes:

Example 651

6.11 Disallowed Raw HTML (extension)

GFM enables the tagfilter extension, where the following HTML tags will be filtered when rendering HTML output:

- <title>
- <textarea>
- <style>
- <xmp>
- <iframe>
- <noembed>

- <noframes>
- <script>
- <plaintext>

Filtering is done by replacing the leading < with the entity <. These tags are chosen in particular as they change how HTML is interpreted in a way unique to them (i.e. nested HTML is interpreted differently), and this is usually undesireable in the context of other rendered Markdown content.

All other HTML tags are left untouched.

Example 653

6.12 Hard line breaks

A line break (not in a code span or HTML tag) that is preceded by two or more spaces and does not occur at the end of a block is parsed as a $\frac{\text{hard line break}}{\text{crendered in HTML as a}}$ (rendered in HTML as a

Example 654

For a more visible alternative, a backslash before the <u>line ending</u> may be used instead of two spaces:

Example 655

More than two spaces can be used:

Example 656

Leading spaces at the beginning of the next line are ignored:

Example 657

foo\	foo
···· bar	bar

Line breaks can occur inside emphasis, links, and other constructs that allow inline content:

Example 659

*foo ••	foo
bar*	bar

Example 660

*foo\	foo
bar*	bar

Line breaks do not occur inside code spans

Example 661

Example 662

`code\ span`	<code>code\ span</code>
Spair	

or HTML tags:

Example 663

Example 664

<a foo\<="" href="foo\</th><th>	
bar">	bar">

Hard line breaks are for separating inline content within a block. Neither syntax for hard line breaks works at the end of a paragraph or other block element:

Example 665

foo\	foo\
Example 666	
foo ••	foo

### foo\	<h3>foo\</h3>
Example 668	
### £00 ·•	<h3>foo</h3>

6.13 Soft line breaks

A regular line break (not in a code span or HTML tag) that is not preceded by two or more spaces or a backslash is parsed as a **softbreak**. (A softbreak may be rendered in HTML either as a <u>line ending</u> or as a space. The result will be the same in browsers. In the examples here, a <u>line ending</u> will be used.)

Example 669



Spaces at the end of the line and beginning of the next line are removed:

Example 670



A conforming parser may render a soft line break in HTML either as a line break or as a space.

A renderer may also provide an option to render soft line breaks as hard line breaks.

6.14 Textual content

Any characters not given an interpretation by the above rules will be parsed as plain textual content.

Example 671

hello \$.; 'there	hello \$.; 'there
Example 672	
Foo «x ρῆ v	Foo «x ρ η̃ v

Internal spaces are preserved verbatim:

Example 673

Multiple •••• spaces	Multiple •••• spaces
----------------------	----------------------

Appendix: A parsing strategy

In this appendix we describe some features of the parsing strategy used in the CommonMark reference implementations.

Overview

Parsing has two phases:

- 1. In the first phase, lines of input are consumed and the block structure of the document—its division into paragraphs, block quotes, list items, and so on—is constructed. Text is assigned to these blocks but not parsed. Link reference definitions are parsed and a map of links is constructed.
- 2. In the second phase, the raw text contents of paragraphs and headings are parsed into sequences of Markdown inline elements (strings, code spans, links, emphasis, and so on), using the map of link references constructed in phase 1.

At each point in processing, the document is represented as a tree of **blocks**. The root of the tree is a document block. The document may have any number of other blocks as **children**. These children may, in turn, have other blocks as **children**. The last child of a block is normally considered **open**, meaning that subsequent lines of input can alter its contents. (Blocks that are not open are **closed**.) Here, for example, is a possible document tree, with the open blocks marked by arrows:

```
-> document
-> block_quote
    paragraph
        "Lorem ipsum dolor\nsit amet."
-> list (type=bullet tight=true bullet_char=-)
        list_item
        paragraph
        "Qui *quodsi iracundia*"
-> list_item
        -> paragraph
        "aliquando id"
```

Phase 1: block structure

Each line that is processed has an effect on this tree. The line is analyzed and, depending on its contents, the document may be altered in one or more of the following ways:

- 1. One or more open blocks may be closed.
- 2. One or more new blocks may be created as children of the last open block.
- 3. Text may be added to the last (deepest) open block remaining on the tree.

Once a line has been incorporated into the tree in this way, it can be discarded, so input can be read in a stream.

For each line, we follow this procedure:

1. First we iterate through the open blocks, starting with the root document, and descending through last children down to the last open block. Each block

imposes a condition that the line must satisfy if the block is to remain open. For example, a block quote requires a ≥ character. A paragraph requires a non-blank line. In this phase we may match all or just some of the open blocks. But we cannot close unmatched blocks yet, because we may have a lazy continuation line.

- 2. Next, after consuming the continuation markers for existing blocks, we look for new block starts (e.g. ≥ for a block quote). If we encounter a new block start, we close any blocks unmatched in step 1 before creating the new block as a child of the last matched block.
- 3. Finally, we look at the remainder of the line (after block markers like ≥, list markers, and indentation have been consumed). This is text that can be incorporated into the last open block (a paragraph, code block, heading, or raw HTML).

Setext headings are formed when we see a line of a paragraph that is a <u>setext</u> <u>heading underline</u>.

Reference link definitions are detected when a paragraph is closed; the accumulated text lines are parsed to see if they begin with one or more reference link definitions. Any remainder becomes a normal paragraph.

We can see how this works by considering how the tree above is generated by four lines of Markdown:

```
> Lorem ipsum dolor
sit amet.
> - Qui *quodsi iracundia*
> - aliquando id
```

At the outset, our document model is just

```
-> document
```

The first line of our text,

```
> Lorem ipsum dolor
```

causes a <code>block_quote</code> block to be created as a child of our open <code>document</code> block, and a <code>paragraph</code> block as a child of the <code>block_quote</code>. Then the text is added to the last open block, the <code>paragraph</code>:

```
-> document

-> block_quote

-> paragraph

"Lorem ipsum dolor"
```

The next line.

```
sit amet.
```

is a "lazy continuation" of the open paragraph, so it gets added to the paragraph's text:

The third line.

```
> - Qui *quodsi iracundia*
```

causes the paragraph block to be closed, and a new list block opened as a child of the block_quote. A list_item is also added as a child of the list, and a paragraph as a child of the list item. The text is then added to the new paragraph:

```
-> document
-> block_quote
    paragraph
        "Lorem ipsum dolor\nsit amet."
-> list (type=bullet tight=true bullet_char=-)
        -> list_item
        -> paragraph
        "Qui *quodsi iracundia*"
```

The fourth line,

```
> - aliquando id
```

causes the $list_item$ (and its child the paragraph) to be closed, and a new $list_item$ opened up as child of the list. A paragraph is added as a child of the new $list_item$, to contain the text. We thus obtain the final tree:

```
-> document
-> block_quote
    paragraph
        "Lorem ipsum dolor\nsit amet."
-> list (type=bullet tight=true bullet_char=-)
        list_item
        paragraph
        "Qui *quodsi iracundia*"
-> list_item
-> paragraph
        "aliquando id"
```

Phase 2: inline structure

Once all of the input has been parsed, all open blocks are closed.

We then "walk the tree," visiting every node, and parse raw string contents of paragraphs and headings as inlines. At this point we have seen all the link reference definitions, so we can resolve reference links as we go.

```
document
  block_quote
  paragraph
    str "Lorem ipsum dolor"
```

```
softbreak
str "sit amet."
list (type=bullet tight=true bullet_char=-)
list_item
paragraph
str "Qui"
emph
str "quodsi iracundia"
list_item
paragraph
str "aliquando id"
```

Notice how the <u>line ending</u> in the first paragraph has been parsed as a <u>softbreak</u>, and the asterisks in the first list item have become an <u>emph</u>.

An algorithm for parsing nested emphasis and links

By far the trickiest part of inline parsing is handling emphasis, strong emphasis, links, and images. This is done using the following algorithm.

When we're parsing inlines and we hit either

```
a run of * or _ characters, ora [ or ! [
```

we insert a text node with these symbols as its literal content, and we add a pointer to this text node to the **delimiter stack**.

The <u>delimiter stack</u> is a doubly linked list. Each element contains a pointer to a text node, plus information about

- the type of delimiter ([, ![, *,])
- the number of delimiters,
- whether the delimiter is "active" (all are active to start), and
- whether the delimiter is a potential opener, a potential closer, or both (which depends on what sort of characters precede and follow the delimiters).

When we hit a] character, we call the *look for link or image* procedure (see below).

When we hit the end of the input, we call the *process emphasis* procedure (see below), with stack_bottom = NULL.

look for link or image

Starting at the top of the delimiter stack, we look backwards through the stack for an opening [or ! [delimiter.

- If we don't find one, we return a literal text node].
- If we do find one, but it's not *active*, we remove the inactive delimiter from the stack, and return a literal text node].
- If we find one and it's active, then we parse ahead to see if we have an inline link/image, reference link/image, compact reference link/image, or shortcut reference link/image.

- If we don't, then we remove the opening delimiter from the delimiter stack and return a literal text node].
- o If we do, then
 - We return a link or image node whose children are the inlines after the text node pointed to by the opening delimiter.
 - We run process emphasis on these inlines, with the [opener as stack bottom.
 - We remove the opening delimiter.
 - If we have a link (and not an image), we also set all [delimiters before the opening delimiter to *inactive*. (This will prevent us from getting links within links.)

process emphasis

Parameter stack_bottom sets a lower bound to how far we descend in the <u>delimiter stack</u>. If it is NULL, we can go all the way to the bottom. Otherwise, we stop before visiting stack bottom.

Let current_position point to the element on the <u>delimiter stack</u> just above stack_bottom (or the first element if stack_bottom is NULL).

We keep track of the openers_bottom for each delimiter type (*, _) and each length of the closing delimiter run (modulo 3). Initialize this to stack_bottom.

Then we repeat the following until we run out of potential closers:

- Move current_position forward in the delimiter stack (if needed) until we find the first potential closer with delimiter * or _. (This will be the potential closer closest to the beginning of the input the first one in parse order.)
- Now, look back in the stack (staying above stack_bottom and the openers_bottom for this delimiter type) for the first matching potential opener ("matching" means same delimiter).
- If one is found:
 - Figure out whether we have emphasis or strong emphasis: if both closer and opener spans have length >= 2, we have strong, otherwise regular.
 - Insert an emph or strong emph node accordingly, after the text node corresponding to the opener.
 - Remove any delimiters between the opener and closer from the delimiter stack.
 - Remove 1 (for regular emph) or 2 (for strong emph) delimiters from the
 opening and closing text nodes. If they become empty as a result,
 remove them and remove the corresponding element of the delimiter
 stack. If the closing node is removed, reset current_position to the
 next element in the stack.

• If none is found:

- Set openers_bottom to the element before current_position. (We know that there are no openers for this kind of closer up to and including this point, so this puts a lower bound on future searches.)
- If the closer at current_position is not a potential opener, remove it from the delimiter stack (since we know it can't be a closer either).
- Advance current_position to the next element in the stack.

After we're done, we remove all delimiters above $\underline{stack_bottom}$ from the delimiter stack.