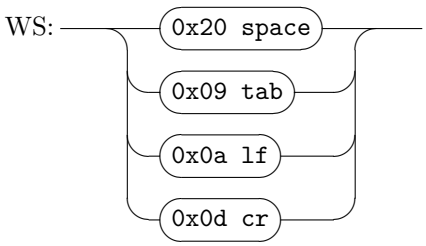
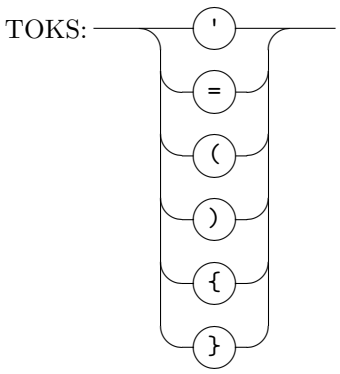
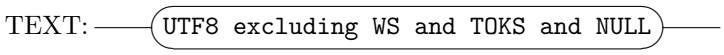
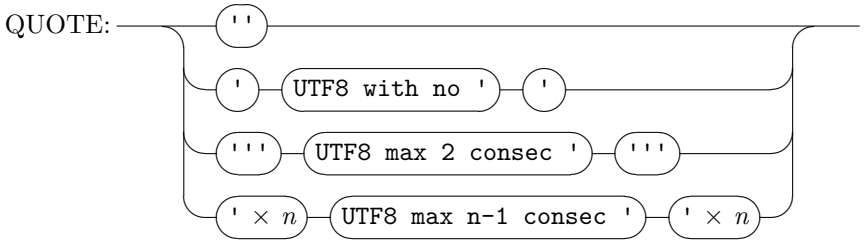
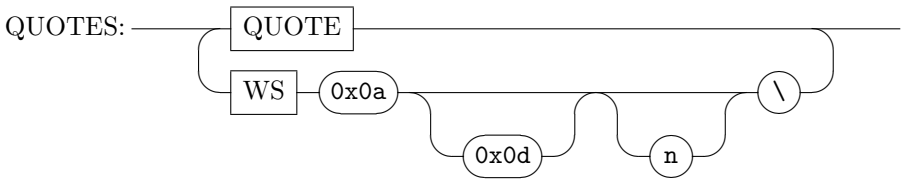
Lexer

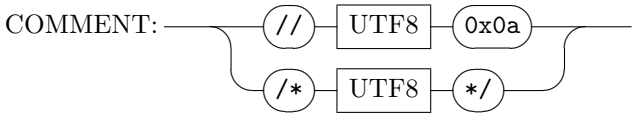

WS:
0x20 space
0x09 tab
0x0a lf
0x0d cr

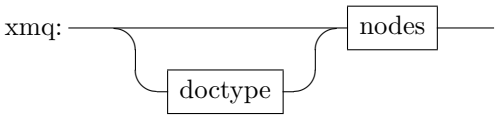

TOKS:
'
=
(
)
{
}

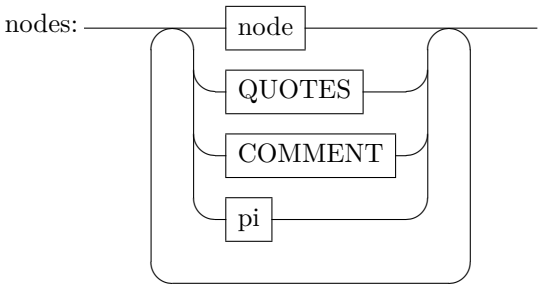

TEXT:
UTF8 excluding WS and TOKS and NULL

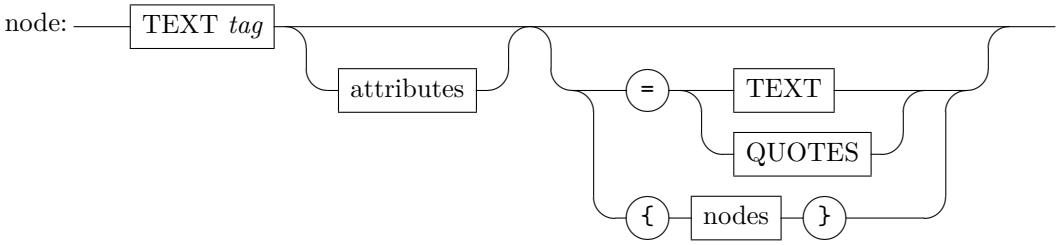

QUOTE:
''
'
UTF8 with no '
'
'''
UTF8 max 2 consec '
'''
' × n
UTF8 max n-1 consec '
' × n

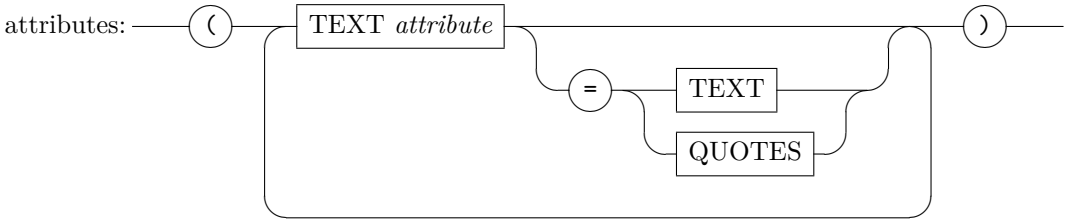

QUOTES:
QUOTE
WS
0x0a
0x0d
n
\

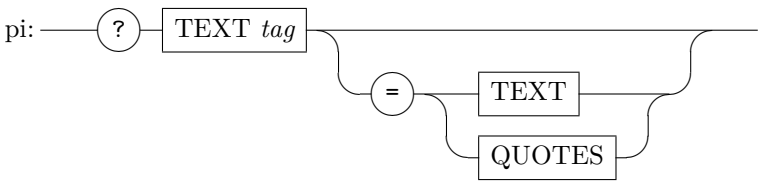

COMMENT:
//
UTF8
0x0a
/*
UTF8
*/

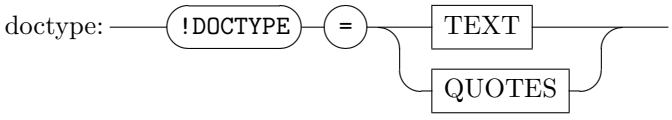

xmq:
nodes
doctype


nodes:
node
QUOTES
COMMENT
pi


node:
TEXT tag
attributes
=
TEXT
QUOTES
{
nodes
}


attributes:
(
TEXT attribute
=
TEXT
QUOTES
)


pi:
?
TEXT tag
=
TEXT
QUOTES


doctype:
!DOCTYPE
=
TEXT
QUOTES

TEXT tag and attribute must conform to the XML element naming rules:

1. must start with a letter or underscore
2. cannot start with the letters xml (or XML, or Xml, etc)
3. can contain letters, digits, hyphens, underscores, and periods

QUOTE content is trimmed according to these rules:

1. Any leading/ending (space|cr|tab)* nl (space|cr|tab)* is removed.
2. Incidental whitespace will always be trimmed, when there is at least one nl left.

QUOTES can combine multiple QUOTE(s):

1. A suffix \ joins the next QUOTE seamlessly.
2. A suffix \n inserts a newline before joining.

Joining children inside the xmq rule:

1. All children are joined seamlessly.
2. Except two consecutive QUOTE are joined with a newline.

Pretty printing introduces whitespace for div{a{span=' hi '}}

1. Default XML output will insert newlines+indentation.

```
<div>
    <a>
        <span> hi </span>
    </a>
</div>
```

2. Default HTML output will prevent newlines+indentation next to known inline elements where whitespace is significant.

```
<div>
    <a><span> hi </span></a></div>
```

3. The option --nopp will prevent all pretty printing and this can be necessary when the DTD/CSS has tailored significant whitespace.

```
<div><a><span> hi </span></a></div>
```