

Initialize:

Set Stream to the input text string
Set currentPosition to 0 and internalQuoteFlag to false
Set delimiterSet to ' , . ; : ! ? () < > + " \n \t space
Set whiteSpace to \t \n space

Procedure getNextToken:

```
L1: cursor := currentPosition; ch := charAt(cursor);
    If ch = endOfStream then return null; endif
L2: while ch is not endOfStream nor instanceOf(delimiterSet) do
    increment cursor by 1; ch := charAt(cursor);
endwhile
If ch = endOfStream then
    If cursor = currentPosition then return null; endif
endif
If ch is whiteSpace then
    If currentPosition = cursor then
        increment currentPosition by 1 and goto L1;
    else
        Token := substring(Stream,currentPosition,cursor-1);
        currentPosition := cursor+1; return Token;
    endif
endif
If ch = ' then
    If charAt(cursor-1) = instanceOf(delimiterSet) then
        internalQuoteFlag := true; increment currentPosition by 1; goto L1;
    endif
    If charAt(cursor+1) != instanceOf(delimiterSet) then
        increment cursor by 1; ch := charAt(cursor); goto L2;
    elseif internalQuoteFlag = true then
        Token := substring(Stream,currentPosition,cursor-1);
        internalQuoteFlag := false;
    else
        Token := substring(Stream,currentPosition,cursor);
    endif
    currentPosition := cursor+1; return Token;
endif
If cursor = currentPosition then
    Token := ch; currentPosition := cursor+1;
else
    Token := substring(Stream,currentPosition,cursor-1);
    currentPosition := cursor;
endif
return Token;
endprocedure
```

Fig. 2.2 Tokenization algorithm

Input: a text with periods

Output: same text with End-of-Sentence (EOS) periods identified

Rules:

All ? ! are EOS

If " or ' appears before period, it is EOS

If the following character is not white space, it is not EOS

If))] before period, it is EOS

If the token to which the period is attached is capitalized
and is < 5 characters and the next token begins uppercase,
it is not EOS

If the token to which the period is attached has other periods,
it is not EOS

If the token to which the period is attached begins with a lowercase
letter and the next token following whitespace is uppercase,
it is EOS

If the token to which the period is attached has < 2 characters,
it is not EOS

If the next token following whitespace begins with \$ ({ [" ' it is EOS
Otherwise, the period is not EOS