Treebank tokenization

Our tokenization is fairly simple:

- most punctuation is split from adjoining words
- double quotes (") are changed to doubled single forward- and backward- quotes (`` and '')
- verb contractions and the Anglo-Saxon genitive of nouns are split into their component morphemes, and each morpheme is tagged separately.
    - Examples
        - `children's --> children 's`
        - `parents' --> parents '`
        - `won't --> wo n't`
        - `gonna --> gon na`
        - `I'm --> I 'm`

  This tokenization allows us to analyze each component separately, so (for example) "I" can be in the subject Noun Phrase while "'m" is the head of the main verb phrase.
- There are some subtleties for hyphens vs. dashes, elipsis dots (...) and so on, but these often depend on the particular corpus or application of the tagged data.
- In parsed corpora, bracket-like characters are converted to special 3-letter sequences, to avoid confusion with parse brackets. Some POS taggers, such as Adwait Ratnaparkhi's MXPOST, require this form for their input.
  In other words, these tokens in POS files: `( ) [ ] { }`
  become, in parsed files: `-LRB- -RRB- -RSB- -RSB- -LCB- -RCB-`
  (The acronyms stand for (Left|Right) (Round|Square|Curly) Bracket.)

Here is a simple sed script that does a decent enough job on most corpora, once the corpus has been formatted into one-sentence-per-line.