A **lemmatizer** delivers the *correct* “dictionary form” of each word (as opposed to a **stemmer**, which simply makes a rough attempt to remove suffixes).

The English inflectional suffixes are `-s` `-ed` `-ing` `-er` `-est`. There are also plenty of irregular forms, such as *eaten*.

To lemmatize, you will need:
- Tagged input (words with POS tags);
- A lexicon of words in the language (could be the same one used for tagging);
- A table of irregular forms (irregular verbs, irregular noun plurals, etc.).

**Example of lemmatization:**
Given *having/VBG*:
- Look in the table of irregular forms; it's not there.
- The general algorithm says you should try to remove `-ing` from anything tagged VBG, provided the result is in the lexicon as VBP or VB.
- The result of removing `-ing` from *having* could be either *have* or *hav*.
- One of these is in the lexicon, so it is used.

By insisting that the lemma be found in the lexicon, we avoid such mistakes as *rabies => raby* (analogous to *babies => baby*).

**Spelling rules** for English are summarized in *Natural Lg. Processing for Prolog Programmers* (and there is one more rule, changing `-c` to `-ck` before a suffix). These spelling rules must be applied when removing a suffix. There is often more than one possibility.

Algorithm on next page...
General lemmatizing algorithm for English (I think):

If the word and tag are in the table of irregular forms, take the lemma from the table.

Else if the word is tagged NNS, NNPS, or VBZ, and removing –s gives you a word that is in the lexicon as NN, NNP, or VBP (or VB) respectively, take that result.

Else if the word is tagged VBG and removing –ing gives you a word that is in the lexicon as VBP or VB, take that result.

Else if the word is tagged VBN or VBD and removing –ed gives you a word that is in the lexicon as VBP or VB, take that result.

Else if the word is tagged JJR or RBR and removing –er gives you a word that is in the lexicon as JJ or RB respectively, take that result.

Else if the word is tagged JJS or RBS and removing –est gives you a word that is in the lexicon as JJ or RB respectively, take that result.

Else leave the word unchanged.