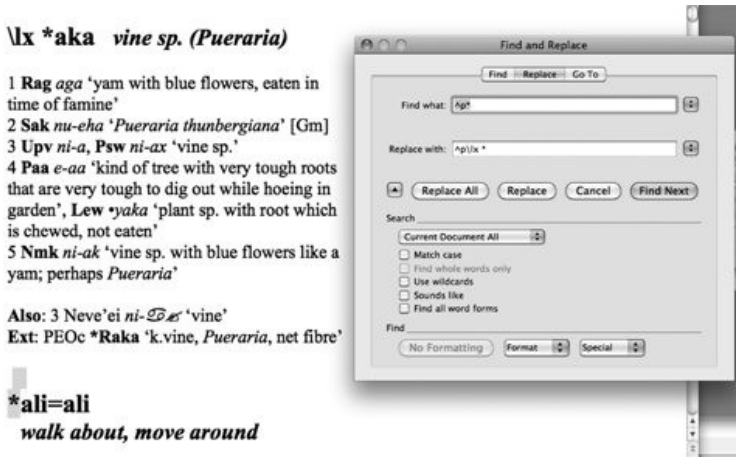
4.8 shows an unstructured Microsoft Word document with a regular pattern of headwords preceded by an
p. 111 asterisk and followed by a definition in italics.[28] Using a regular expression in the ↳

p. 112 ↳

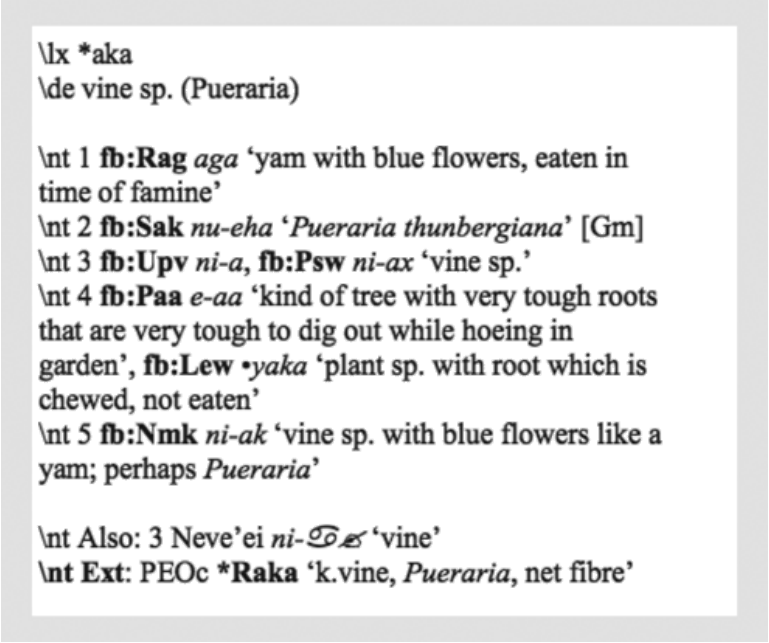
Word find-and-replace window locates a carriage return followed by an asterisk and inserts '\lx' as a field marker identifying the headword, as has been done for **aka*.

Figure 4.8.



Limited regular expression search in MS Word, inserting '\lx' before each headword in a document.

Figure 4.9.



Second insertion of codes into a document on the way to structuring all elements.

Eventually, more codes can be inserted, as shown in Fig. 4.9, where the '\de' and '\nt' fields have been added. Ultimately, all elements of the entries will be explicitly coded, rather than relying on formatting to imply structure.

As a third example of the use of regular expressions, imagine you have a corpus for which you need to quantify the occurrence of a particular word token expressed as a proportion of the total number of tokens