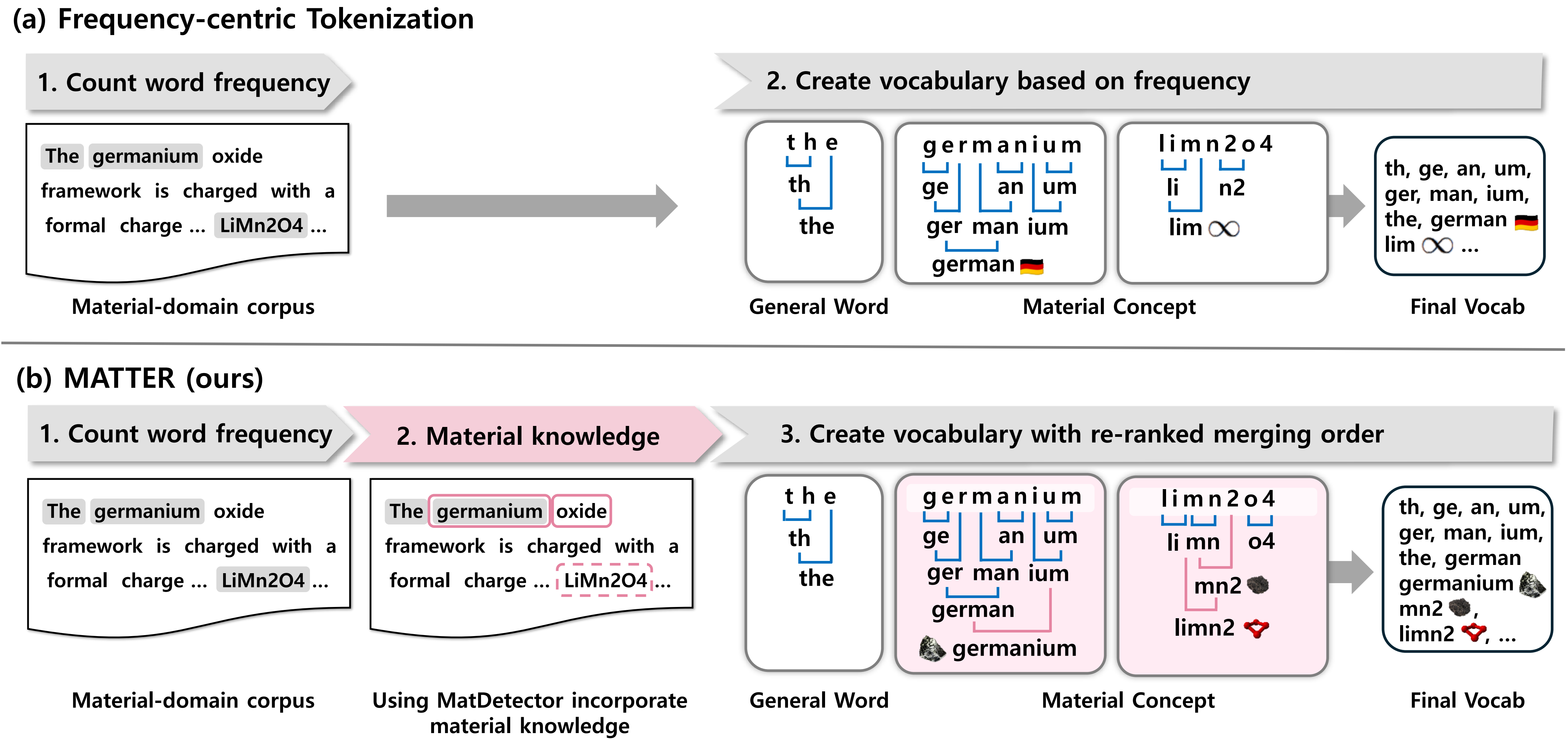
(a) Frequency-centric Tokenization
1. Count word frequency
The germanium oxide framework is charged with a formal charge ... LiMn2O4 ...
Material-domain corpus
2. Create vocabulary based on frequency
t h e
th
the
General Word
g e r m a n i u m
ge an um
ger man ium
german
l i m n 2 o 4
li n2
lim
Material Concept
th, ge, an, um, ger, man, ium, the, german lim ...
Final Vocab
(b) MATTER (ours)
1. Count word frequency
2. Material knowledge
3. Create vocabulary with re-ranked merging order
The germanium oxide framework is charged with a formal charge ... LiMn2O4 ...
Material-domain corpus
The germanium oxide framework is charged with a formal charge ... LiMn2O4 ...
Using MatDetector incorporate material knowledge
t h e
th
the
General Word
g e r m a n i u m
ge an um
ger man ium
german
germanium
l i m n 2 o 4
li mn o4
mn2
limn2
Material Concept
th, ge, an, um, ger, man, ium, the, german germanium mn2 , limn2 , ...
Final Vocab