```
## Examples

txt <- "The United States ended its fiscal year with funding for
        social security and health care."

tokens(txt)
# tokens from 1 document.
# text1 :
#  [1] "The"      "United"   "States"   "ended"    "its"      "fiscal"   "year"
#  [8] "with"     "funding"  "for"      "social"   "security" "and"      "health"
# [15] "care"     "."

tokens(txt) %>%
    tokens_compound(head(colls, 10))
# tokens from 1 document.
# text1 :
#  [1] "The"             "United_States"   "ended"           "its"
#  [5] "fiscal_year"     "with"            "funding"         "for"
#  [9] "social_security" "and"             "health_care"     "."
```

```
## Compounding tokens in MWEs (cont.)

##  Knows not to concatenate across punctuation etc boundaries

tokens("One two three, four and five.") %>%
    tokens_compound(list(c("one", "two"), c("three", "four"),
                         c("four", "five")))
# tokens from 1 document.
# text1 :
#  [1] "One_two" "three"   ","       "four"    "and"     "five"    "."
```

```
## Alternatives: Named entity recognition}

sp_ne <- spacy_parse(data_corpus_sotu, entity = TRUE)
ne <- entity_extract(sp_ne)
table(grep("_", ne$entity, value = TRUE)) %>%
    sort(decreasing = TRUE) %>%
    head(20)
#            the_United_States                Great_Britain       the_Federal_Government
#                         4042                          493                          328
#                United_States       the_General_Government            The_United_States
#                          237                          162                          160
# the_House_of_Representatives             the_Soviet_Union     the_Postmaster_-_General
#                          155                          150                          143
#     the_District_of_Columbia                     New_York              Social_Security
#                          135                          122                          114
#           the_United_Nations       the_British_Government      the_National_Government
#                          114                          105                          105
#            the_Supreme_Court              the_Middle_East      the_Department_of_State
#                          105                          103                          100
#     House_of_Representatives              Central_America
#                           92                           88
  \end{verbatim}
 \end{frame}
```

Alternatives: Noun phrase recognition

```
sp_np <- spacy_parse(data_corpus_sotu, nounphrase = TRUE)
np <- nounphrase_extract(sp_np)
table(grep("_\\w+_", np$nounphrase, value = TRUE)) %>%
    sort(decreasing = TRUE) %>%
    head(20)
#      the_United_States    the_American_people        the_fiscal_year
#                   3345                    300                    265
#          the_same_time the_Federal_Government       the_last_session
#                    261                    256                    237
#          the_past_year        the_public_debt      the_two_countries
#                    215                    176                    148
# the_General_Government      The_United_States     the_public_service
#                    135                    127                    123
#         the_first_time       the_public_lands       the_Soviet_Union
#                    121                    118                    116
#       the_present_year   the_last_fiscal_year          the_last_year
#                    106                    105                    104
#    the_two_Governments     the_several_States
#                    102                     98
```