WHAT IS A WORD?

A sentence is a string of words.

So we need to start with a topic which seems to be obvious but isn’t when looked at in more detail.

***tokenization*: An early step of sentence processing is to divide the input text into *tokens*, where each token is a word or something like it, e.g. a number or punctuation mark.**

Periods are necessary to know where sentences end.

Commas and other punctuation are useful to tell us where phrases end.

First suggestion: word = “string of contiguous alphanumeric characters with space on either side, including hyphens and apostrophes but no other punctuation”.

Problems with this suggestion:

> **$22.50**
>
> **Micro$oft**
>
> **C|net (a web company)**
>
> **:)**

Second suggestion: word = string separated by whitespace

Problems with second suggestion:

> **Any word that ends a sentence is a problem.**

etc.	**[doesn’t end a sentence]**
Calif.	**[same here]**
Wash.	**[and here]**
Wash.	**[but here it does! (from washing machine instructions)]**

> **etc. at the end of a sentence = the period serves 2 functions, so we can’t just drop it.**

Quote marks cause problems too.

the boys’ toys	**[apostrophe seems like part of the word here]**
‘Boys’ is a noun.	**[but not here]**

Does ’s mark a word?

> **The Queen of England’s hat has flowers.**

Does this make “Queen of England” a word?

What about hyphens?

These look like one word:
e-mail
co-operate
A-1-plus commercial paper

What about these?
non-lawyer
pro-democracy
so-called

These are even further apart:
once-quiet cars
aluminum-export ban
text-based medium

And these:
a take-it-or-leave-it offer
a 90-cent-an-hour raise
a 22-year-old
the 22-to-45 age group

How to represent that different spellings are the “same word”?
e-mail vs. email
markup vs. mark-up
data base vs. database

Other languages have the same problems, and some different ones too.
German: Lebensversicherungsgesellschaftangestellter =
life insurance company employee

leben = life
versicherung = insurance (ver + sicher +ung)
gesellschaft = company
angestellter = employee

Above, we had no white space, but wanted a word break.
We can also have white space, but don’t want a word break.
815-753-6944
the New York-New Haven railroad [maybe York-New is a word?]

in spite of
in order to
because of

to work out [phrasal verbs]
I couldn’t work the answer out. [...are not always contiguous]

She worked out of her house. [not the phrasal verb here!]
He climbed up the mountain. [is this a phrasal verb?]

WHAT IS A SENTENCE?

About 90% of periods are sentence boundary indicators.

What to do here:

> “You remind me,” she remarked, “of your mother.”

Note that the sentence-ending period is inside the quote marks (North American convention).

Need a good algorithm for sentence identification to do sentence alignment (matching up sentences in multiple languages for machine translation).