Tokenization

Tokenization

Input: raw text

Output: sequence of **tokens** normalized for easier processing.

Tokenization

- Some Asian languages have obvious issues:

  利比亚“全国过渡委员会”执行委员会主席凯卜22日在首都的黎波里公布“过渡政府”内阁名单，宣告过渡政府正式成立。

- But German too: Noun-noun compounds:

  Gesundheitsversicherungsgesellschaften

- Spanish clitics: *Darmelo*
- Even English has issues, to a small degree: *Gregg and Bob’s house*

Tokenization

- Some Asian languages have obvious issues:

  利比亚“全国过渡委员会”执行委员会主席凯卜22日在首都的黎波里公布“过渡政府”内阁名单，宣告过渡政府正式成立。

- But German too: Noun-noun compounds:

  Gesundheits-versicherungs-gesellschaften (health insurance companies)

- Spanish clitics: *Darmelo*
- Even English has issues, to a small degree: *Gregg and Bob’s house*

Tokenization

- Some Asian languages have obvious issues:

  利比亚“全国过渡委员会”执行委员会主席凯卜22日在首都的黎波里公布“过渡政府”内阁名单，宣告过渡政府正式成立。

- But German too: Noun-noun compounds:

  Gesundheitsversicherungsgesellschaften

- Spanish clitics: *Dar-me-lo (To give me it)*
- Even English has issues, to a smaller degree: *Gregg and Bob’s house*

Tokenization

Input: raw text

```
Dr. Smith said tokenization of English is “harder than you’ve thought.”
When in New York, he paid $12.00 a day for lunch and wondered what it would
be like to work for AT&T or Google, Inc.
```

Output from Stanford Parser: http://nlp.stanford.edu:8080/parser/index.jsp
with part-of-speech tags:

```
Dr./NNP Smith/NNP said/VBD tokenization/NN of/IN English/NNP
is/VBZ ``/`` harder/JJR than/IN you/PRP 've/VBP thought/VBN ./.
''/’’

When/WRB in/IN New/NNP York/NNP ,/, he/PRP paid/VBD $/$ 12.00/CD
a/DT day/NN for/IN lunch/NN and/CC wondered/VBD what/WP it/PRP
would/MD be/VB like/JJ to/TO work/VB for/IN AT&T/NNP or/CC
Google/NNP ,/, Inc./NNP ./.
```