"Latest figures from the US government show the trade deficit with China reached an all-time high of $365.7bn (£250.1bn) last year. By February this year it had already reached $57bn."

"Latest figures from the US government show the trade deficit with China reached an all time high of $ 365.7 bn ( £ 250.1 bn ) last year . By February this year it had already reached $ 57 bn ."
Tokenization

- Issues related to tokenization:
  - Separators: punctuations
    - Exceptions: „m.p.h“, „Ph.D“
  - Expansions: „we're“ = „we are“
  - Multi-words expressions: „New York“, „doghouse“
Segmentation = Tokenization

- Word segmentation: separation of the morphemes but also tokenization for languages without 'space' character

Where are the words?
Segmentation?

Improve production uptime and efficiency, while lowering maintenance costs
Sentence separation (splitting)

- Also usually based on punctuations (.,?!,)
  - Exceptions: „Mr.“, „4.5“
Approaches for tokenization

- Based on regular expressions

- Based on rules or machine learning
  - Binary classifiers that decides whether a certain punctuation is part of a word or not
Approaches for segmentation

• Maximum matching approach
  – Based on a dictionary
  – Longest sequence of letters that forms a word

• Palmer (2000):
  \[
  \text{thetabledownthere}
  \]
  \[
  \text{thetabledownthere}
  \]
  \[
  \text{thetabledownthere}
  \]
  \[
  \text{thetabledownthere}
  \]
Neural networks for segmentation

- Chinese word segmentation formalized as a **character-based sequence labeling task** where only contextual information within fixed sized local windows and simple interactions between adjacent tags can be captured.

Figure 1: Our framework.

(https://arxiv.org/abs/1606.04300)
Neural networks for segmentation

(https://arxiv.org/abs/1606.04300)
Tools for tokenization

- Spacy: https://spacy.io/
- OpenNLP: https://opennlp.apache.org/
- Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP/
- deeplearning4j tokenizer: https://deeplearning4j.org/tokenization
- Neural tokenizer: https://github.com/Kyubyong/neural_tokenizer
Exercise

- Project: choose a tokenizer and try it in your document collection.
  - Manually check a sample of the results.
Further reading

- NLP book: Chapter 3