LING-L 445: Computation and Linguistic Analysis
Practical 01B

Instructor: Francis M. Tyers

Dante Razo

February 11, 2019

Segmentation

How should you segment sentences with semicolons? As a single sentence or as two sentences? Should it depend on context?

Should sentence with ellipsis (...) be treated as a single sentence or as several sentences?

If there is an exclamation after the first word in a sentence should it be a separate sentence? How about if there is a comma?

Can you think of some hard tasks for the segmenter?

Tokenization

Why should we split punctuation from the token it goes with?

Should abbreviations with space in them be written as a single token or two tokens?

How about numerals like 134 000?

If you have a case suffix following punctuation, how should it be tokenized ?

Should contractions and clitics be a single token or two (or more) tokens?