

## General introduction

- ▶ What the course is about
- ▶ How we do this course (syllabus)
- ▶ A few basic linguistic concepts
- ▶ A few basic machine-learning concepts

## Tokens, types, and lemmas

*I rose to saw off the still rose that I saw still grew near the still.*

16 tokens.

Types occurring once: to, off, that, grew, near.

Types occurring twice: I, rose, saw, the.

Type occurring three times: still

Distinct lemmas (lexemes): I, rise, to, saw (verb-cut), off, the, still (adjective), rose (noun), that, see, still (adverb), grow, near, still (noun).

For next time: **Learn Python**, take quiz

Coming up:

- ▶ Trying out Python and NLTK (Friday)
- ▶ Regular expressions (Monday)
- ▶ Fun with regular expressions (Wednesday)
- ▶ Starting language models (next week Friday)