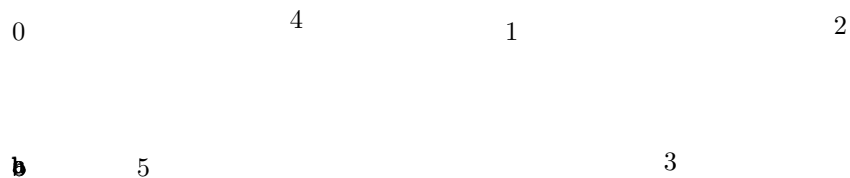**Exercise sheet 1**

(Submit as a plain text email message with subject "Homework 1" to dm@ling.osu.edu on
Tuesday, Jan. 10.)

1. Consider the following finite-state machine:

   0   4   1   2

   b   5   3

   (a) Which of the following sequences does it accept? (1) *ab* (2) *ca* (3) *cba* (4) *cabbb* (5) *ababa* (6) *bacb* (7) *cabcb* (8) *ababcbc* (9) ϵ (10) *cabbbab*

   (b) Write a regular expression which characterizes the same language as this network.

2. Go to the site http://www.lexmasterclass.com/exercises/regex/index.html and do the exercises 1 to 4, sending me the four regular expressions.

3. Specify a regular expression that recognizes English number names up to (and including) thirty (e.g., *one*, *two*, *fourteen*, *twenty-two*) in the various possible spellings.

   You can find the BNC sampler corpus, with one word per line, on the linguistics department unix machines (e.g., on julius.ling.ohio-state.edu), in the file /home/scratch/dm/bnc-samp.txt. Beware, this is a large file (2,427,451 tokens), so do **not** copy it into your home directory!

   Use `egrep` to test (and refine) your regular expression on this file and report both the regular expression you end up with and how many English numbers up to (and including) thirty are found by it.

   Example:

   `egrep -c '^(one|two)$'` /home/scratch/dm/bnc-samp.txt finds 11343 occurrences.

   You might find Stuart Robinson: "Grep for Linguists" useful, which you can find at:

   http://arts.anu.edu.au/linguistics/misc/comp_resources/grep.html