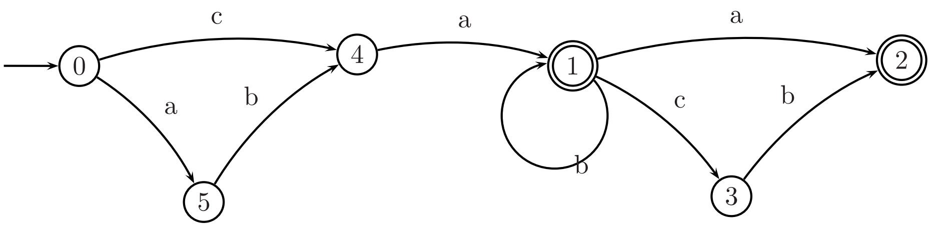
Introduction to CL I (684.01)
Detmar Meurers

OSU Linguistics
Winter 2004

Exercise sheet 1

(Submit by email to dm@ling.osu.edu before class on Tuesday, 13. January)

1. Consider the following finite-state machine:



   (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.itri.brighton.ac.uk/courses/MScLex/exercises/regex/` 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 thirty (e.g., *one*, *two*, *fourteen*) in the various possible spellings.

   You can find the BNC sampler corpus, with one word per line, 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 twenty are found by it.

   Example:

   `egrep -c ’^(one|two)$’ /home/scratch/dm/bnc-samp.txt` finds 12036 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`