Language and Computation

LING 227 01 / 627 01 / PSYC 327 01
Assignment #1
Due: February 4, 2014

The solutions to the problem set must be handed in on paper at the beginning of the class. It must be typed (printed), including mathematical formulae. (Graphs of finite-state automata, etc. may be added by hand.)

The goal of this homework is manifold: it should help you practice certain concepts, deepen your understanding of the material, test your understanding of the textbook, but also prepare you for topics and notions to be introduced later in the course. Please be advised that I prefer a teaching style that intertwines the learning process with measuring its results (that is, even the final exam is part and parcel of the learning process).

Each problem set will be worth 10 points in total.

The problem set is based on the first three chapters of Jurafsky & Martin. Finite-state transducers will be amply covered in details in the lecture on 01/28, a week before the deadline.

Problem 1: FSA as a realistic model? (2 points)

Many aspects of language have been approached using finite-state technology. You can search the web for *FSMNLP*, the workshop series on Finite-State Methods and Natural Language Processing, and check `http://www.aclweb.org/aclwiki/index.php?title=SIGFSM` for SIGFSM, the *ACL Special Interest Group on Finite-State Methods.*

Yet, what is the relevance of this machinery for the study of language? On the one hand, can it contribute to our understanding of language as a product of the human mind/brain? On the other, is a finite-state model of language representative to what a computer can make of language?

Present your opinion regarding the following two questions:

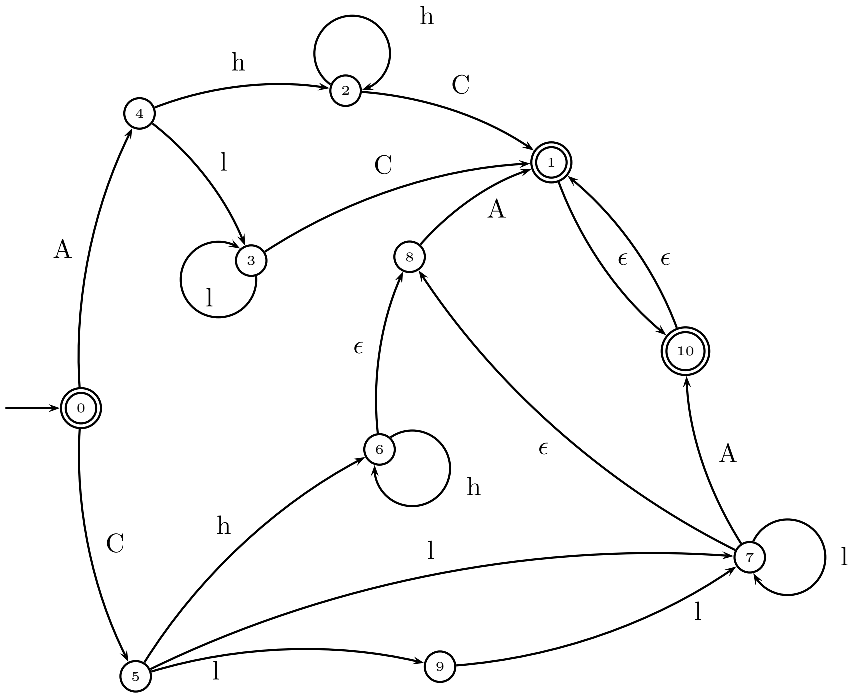
1. Can the human brain be viewed as a finite-state automaton/transducer?
2. Can a computer be viewed as a finite-state automaton/transducer?

Summarize your arguments in a prose of approx. 0.5 to 1 page in length. You need not necessarily argue for either a definitive 'yes' or a definitive 'no'.

Problem 2: Language accepted (2 points)

Our alphabet consists of the following four letters: $\Sigma = \{$A[nne], C[harlie], h[ates], l[oves]$\}$. What is the language accepted by the finite-state automaton below? The start state has an incoming arrow, and the final states are represented with double circles.

1. Use prose to answer this question.
2. Answer this question using regular expressions.
3. Can you simplify the automaton without changing the language accepted?



(Graph made using the export-to-LaTeX function of Gertjan van Noord's `FSA Utilities`.)

Problem 3: Dates and regular expressions (4 points)

Dates can be written in different ways: *January 23, 2014* or *01/23/2014*, let alone *23/01/2014* and formats using dashes or dots instead of slashes (refer to, e.g., `http://en.wikipedia.org/wiki/Calendar_date#Date_format`).

Part 1: Write a regular expression that only matches strings that can be dates in the Gregorian calendar using the `MM/DD/YYYY` format. Several solutions are possible. Use the regex syntax introduced in JM Chapter 2. You may want to balance between a perfect solution and a short solution: why?

Part 2: Draw a finite-state automaton equivalent to your regular expression just developed. (If you have presented more than one solutions in Part 1, then you can choose a reasonably complex one.) Explain shortly what each state "stands for": when during the recognition of a date the FSA is in that state.

Part 3: Write an "extended" regular expression (using *registers*) that transforms ("substitutes") dates in `MM-DD-YYYY` format to `DD.MM.YYYY` format.

Part 4: Describe (no need to fully specify) a transducer (either a regex or an FST) that transforms dates in `MM/DD/YYYY` format into `Month DD, YYYY` format (such as *January 23, 2014*).

Problem 4: Finite-state transducers as a tool for machine translation? (2 points)

Imagine you have to implement a machine translation system, but the only machinery you can use is a finite-state transducer. What will you be able to do, and what won't? Discuss that question in a prose of approx. 0.5 to 1 page in length.[1]

Think of different levels and aspects of language: syntax, morphology, phonology, orthography, etc. Bring examples from languages you speak and/or from imaginary languages. What language pairs would best suit the task of "finite-state machine translation"?

[1] Note that untrained translators, especially in the past, often almost acted as a finite-state transducer. Any such experience?