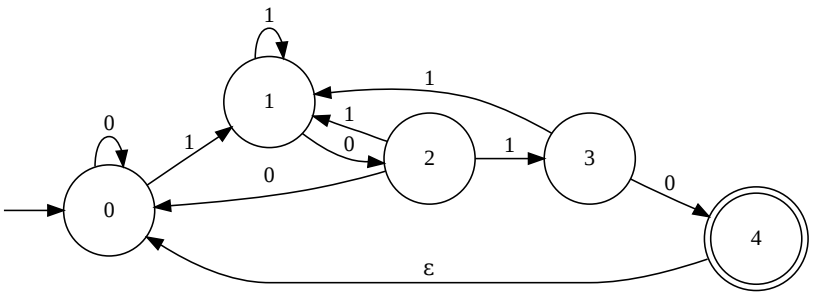
Einführung in die Computerlinguistik
Manfred Pinkal **WS 2008/09**

Übungsblatt 4, Abgabedatum 24.11.2008

Die Übungsblätter können in Gruppen von bis zu drei Personen bearbeitet werden. Bitte denken Sie daran ihren Namen auf das Blatt zu schreiben. Bei Abgabe per e-Mail bitte die Lösungen ins PDF Format konvertieren.

1. Betrachten Sie den folgenden NEA A_1 über dem Alphabet $\Sigma = \{1, 0\}$:



   a) Beschreiben Sie die Sprache, die der Automat erkennt.
   Bonus: Der Automat erkennt eine bestimmte Teilmenge der natürlichen Zahlen in Binärschreibweise. Um welche Menge handelt es sich?
   b) Konstruieren Sie einen zu A_1 äquivalenten buchstabierenden Automaten.
   c) Erzeugen Sie durch Potenzautomatenkonstruktion einen DEA, der die gleiche Sprache wie A_1 erkennt.

2. a) Lesen Sie im Handbuch Carstensen et al. den Abschnitt 5.1 (Korrekturprogramme), und zwar bis einschließlich 5.1.2 (Kontextabhängige Korrektur).
   b) Eine ältere Version der MS Word-Rechtschreibprüfung hat mir vor Jahren als Alternativen für "semantisch" die Wörter "seemännisch" und "romantisch" angegeben. Bestimmen Sie die im Text genannte Levenshtein-Distanz zwischen den drei Wörtern (also auch zwischen "seemännisch" und "romantisch").
   c) Sagen Sie in kurzen Worten, wozu man bei der Rechtschreibprüfung kontextabhängige Korrektur benötigt.

3. Auf der Webseite `http://community.languagetool.org/` können Sie eine Software für regelbasierte Grammatikprüfung testen.

   Die Startseite präsentiert Ihnen drei zufällig ausgewählte Sätze in denen die Regeln Fehler gefunden haben. Oft erkennt das System falsche Positive (korrekte Konstruktionen werden als Fehler markiert). Betrachten Sie drei interessante englische Beispiele für solche Fälle [1] und die jeweils angewendete Regel [2] [3].

   a) Warum funktioniert die Regel nicht wie beabsichtigt?
   b) Wie ließe sich die Regel so abändern, dass das jeweilige Beispiel korrekt behandelt wird?

[1] Mit 'Show other random examples' können Sie neue Beispiele erzeugen
[2] Klicken Sie dazu auf den Link "[Visit Rule]"
[3] Die meisten Regeln basieren auf Mustern von Wortarten. Die Software verwendet das Penn Tagset, eine Übersicht finden Sie z.B. auf `http://www.computing.dcu.ie/~acahill/tagset.html`. Zusätzlich sind folgende Tags definiert: `NN:U - Mass noun` und `NN:UN - Noun used as mass`