4. Übungsblatt - Abgabe: 22.11.2010

Aufgabe 4.1

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). Finden Sie drei möglichst unterschiedliche englische Beispiele für solche falsch positiven Fälle, bei denen die Regeln auf der Abfolge der Wörter basieren (nicht auf Zeichensetzungs- oder Formatierungsfehlern)[1] und geben Sie die jeweils angewendete Regel an.[2][3] Beantworten Sie für jedes Ihrer 3 Beispiele die folgenden Fragen:

(a) Warum funktioniert die Regel nicht wie beabsichtigt?

(b) Ließe sich die Regel so abändern, dass das jeweilige Beispiel korrekt behandelt wird? Wenn ja, wie? Wenn nein, warum nicht?

Aufgabe 4.2

(a) Schreiben Sie eine kontextfreie Grammatik, die Nominalphrasen wie die folgenden erzeugt:

> **Det N** (*das Auto*)
> **Det A N** (*das neue schnelle Auto*)
> **Det A N Prp Det N** (*das grüne Auto auf dem Parkplatz*)
> **Det N Prp Det N Prp Det A N** (*das Auto auf dem Parkpatz bei dem neuen Institutsgebäude*)
> **Det N Prp PN** (*das Auto von Peter*) **Det N Prp Pro** (*das Auto von ihm*)

Verwenden Sie zusätzliche Kategoriensymbole (z.B. PP für Präpositionalphrase und AP für Adjektivphrase). Schreiben Sie außerdem einige lexikalische Einträge für jede lexikalische Kategorie.

(b) Fügen Sie die NP-Regeln aus (a) zur Grammatik G1 aus den aktuellen Vorlesungsfolien hinzu und leiten Sie drei unterschiedliche Sätze ab (bitte mit den zugehörigen Ableitungsbäumen; der komplette Ableitungsprozess braucht nicht aufgeschrieben zu werden). Mindestens zwei der drei Sätze sollen ziemlich lang sein ($\geq$ 10 Wörter); bitte strukturell möglichst unterschiedliche Sätze ableiten.

[1] Mit 'Show other random examples' können Sie neue Beispiele erzeugen.
[2] Klicken Sie dazu auf den Link "[Visit Rule]".
[3] 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`

(c) Gibt es mit der Grammatik Probleme? Ableitbare Ketten, die keine grammatischen Sätze des Deutschen sind; grammatische Sätze, die eigentlich in den Bereich der Grammatik fallen sollten, aber nicht von ihr erzeugt werden? Bitte geben Sie jeweils ein illustrierendes Beispiel dazu!

Aufgabe 4.3

Das Stuttgart-Tübinger Tagset (STTS) wurde dazu entwickelt, deutsche Texte mit feinkörnigen Wortartinformationen zu annotieren. Bestimmen Sie für die folgenden Sätze die Wortarten nach dem STTS (vgl. das Handout aus der Übung oder http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html). Markieren Sie außerdem alle NPs (können geschachtelt sein). Welche werden von Ihren Regeln aus Aufgabe 4.2 erkannt (das Vorhandensein entsprechender lexikalischer Einträge vorausgesetzt), welche nicht? Abstrahieren Sie davon, dass bestimmte Kategorien im STTS anders als in unserer Grammatik benannt sind (z.B. Det vs. ART).

(a) Die Universität des Saarlandes wurde 1948 mit französischer Unterstützung in dem damals politisch teilautonomen und wirtschaftlich mit Frankreich verbundenen Saarland gegründet.

(b) „Hm“, sagte Peter, „ich würde ja doch gerne wissen, wo ich meinen Schirm liegen gelassen habe oder ob ihn mir irgendjemand geklaut hat.“

Aufgabe 4.4

Entscheiden Sie für die folgenden Sprachen, ob sie regulär oder kontextfrei sind.

(a) $L_1 = \{wcw^R \mid w \in \{a,b\}^*\}$[4]

(b) $L_2 = \{w | w \in \{0,1,2,3,4,5,6,7,8,9\}^*$ und w ist durch 8 teilbar$\}$

(c) $L_3 = \{a^*b^mcb^ma | m \in \mathbb{N}, m \geq 42\}$

(d) $L_4 = \{a^mbbbc^{2m}c^* | m \in \mathbb{N}, m \leq 5\}$

Für den Fall, dass die jeweilige Sprache regulär ist, begründen Sie kurz ihre Entscheidung. Andernfalls argumentieren sie mit dem Pumping Lemma, warum die Sprache nicht regulär ist und zeigen Sie, dass die Sprache kontextfrei ist.

Abgabe in Gruppen von bis zu drei Studierenden bis **22.11.2010** 18 Uhr entweder als Email im pdf-Format an **i2cl@coli.uni-sb.de** oder auf Papier im Briefkasten an der Tür von Raum 1.04 in C7.2.

[4]Erläuterung: w^R ist die Spiegelung von w, d.h. es enthält die Zeichen von w in umgekehrter Reihenfolge. Worte von L_1 sind also z.B. *c*, *abcba*, *bbbaabacabaabbb*