Dr. Vera Demberg
Andrea Horbach

4. Übungsblatt - Abgabe: 25.11.2014

Aufgabe 4.1

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

Die Seite *http://tools.wmflabs.org/languagetool/feedMatches/list?lang=en* präsentiert Ihnen Beispiele aus der englischen Wikipedia, in denen die Regeln Fehler gefunden haben. Oft erkennt das System falsche Positive (korrekte Konstruktionen werden als Fehler markiert). Finden Sie zwei möglichst unterschiedliche englische Beispiele für solche falsch positiven Fälle, bei denen die Regeln auf der Abfolge bzw dem Numerus der Wörter basieren (nicht auf Zeichensetzungs- oder Formatierungsfehlern, geben Sie als Kategorie, nach der gefiltert wird *grammar* an.) und geben Sie die jeweils angewendete Regel an. [1] Beantworten Sie für jedes Ihrer 2 Beispiele die folgenden Fragen:

(a) Was macht die Regel? Wenn Sie die Regel anklicken, können Sie sich das XML-Muster, das hinter der Regel steht anschauen. Versuchen Sie, das Muster zu verstehen und mit Ihren Worten zu beschreiben, wie die Regel funktioniert.

(b) Warum funktioniert die Regel in dem gefundenen Beispiel nicht wie beabsichtigt?

(c) 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.

[1] 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`

(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.

(c) Gibt es mit der Grammatik Probleme? Ableitbare Ketten, für die es keine grammatischen Sätze gibt (Wenn es **auch** nichtgrammatische Instanziierungen einer ableitbaren Kette gibt, ist das kein Problem. Problematisch ist nur der Fall, wenn **alle** möglichen Instanziierungen ungrammatisch sind.); grammatische Sätze, die eigentlich in den Bereich der Grammatik fallen sollten, aber nicht von ihr erzeugt werden? Bitte geben Sie gegebenenfalls 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. http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/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 Computerlinguistik in Saarbrücken ist in einer Forschungstradition, die auf die frühen Siebziger zurückgeht, verwurzelt. Die hochgradig interdisziplinäre Arbeit begann zwischen 1973 und 1986 im Rahmen des SFB 100.

(b) „Ach, wenn ich das gewusst hätte, dann hätte ich uns für heute Abend zwei Kinokarten besorgt, damit wir bei diesem schlechten Wetter irgendwo im Trockenen sitzen können.“

Aufgabe 4.4

Auf der Vorlesungsseite ist eine Datei mit einem Ausschnitt des Tiger-Corpus verlinkt (tiger_pos.txt), der zeilenweise ein Token und seinen POS-Tag enthält. Ihre Aufgabe ist es, auszuzählen, wie häufig die einzelnen POS-Tags in der Datei vorkommen. Dazu sollen Sie die folgenden Befehle nutzen und durch | miteinander kombinieren: *cat*, *cut*, *sort*, *uniq*. Schauen Sie sich an, wie diese Befehle funnktionieren und welche Argumente sie nutzen.

Welchen Gesamt-Befehl haben Sie benutzt? Was ist der häufigste und der seltenste POS-Tag und wie oft kommen sie vor?

Aufgabe 4.5

Entwerfen Sie ein Grammatikfragment für eine Sprache, die nicht Deutsch und wenn möglich auch nicht Englisch ist. Ihr Grammatikfragment sollte mindestens 8 nichtlexikalische Regeln beinhalten (d.h. Regeln, die nicht die Form $ART \rightarrow der$ haben), verschiedene Sätze der Sprache ableiten können und sich im Bereich der Wort-/Satzstellung vom Deutschen unterscheiden. Erklären Sie Ihre vom Deutschen abweichenden Kategoriensymbole. Geben Sie auch lexikalische Einträge an und geben Sie mindestens 3 Beispielableitungen an und erklären Sie diese. Bitte übersetzen Sie Ihre Beispiele so, dass auch jemand, der die Sprache nicht beherrscht, Ihre Beispielableitungen nachvollziehen kann, indem sie zusätzlich zu Ihrer tatsächlichen Übersetzung noch eine Wort-für-Wort-Übersetzung angeben.

Beispiel für eine Wort-für-Wort-Übersetzung:

Je	le	lui	ai	donné.
Ich	es (dirObj)	ihm	habe	gegeben.

Ich habe es ihm gegeben.

Abgabe in Gruppen von bis zu drei Studierenden am **25.11.2014** vor der Vorlesung.