

Aufgabe 1

- (a) Laden Sie den Brown Corpus herunter (<http://www.archive.org/details/BrownCorpus>)
(b) Versuchen Sie aus der Dokumentation (<http://icame.uib.no/brown/bcm.html#bc9>) mehr über das Korpus herauszufinden.

Welche Genres (Textsorten) gibt es, wie sind diese weiter unterteilt.

- (c) Finden Sie für das nachfolgende Fragment aus dem Text cg28 heraus, was die verwendeten Tags bedeuten.

Important/jj as/cs was/bedz Mr./np O'Donnell's/np\$ essay/nn ./, his/pp\$ thesis/nn is/bez so/ql restricting/jj as/cs to/to deny/vb Faulkner/np the/at stature/nn which/wdt he/pps obviously/rb has/hvz ./.

He/pps and/cc also/rb Mr./np Cowley/np and/cc Mr./np Warren/np have/hv fallen/vbn to/in the/at temptation/nn which/wdt besets/vbz many/ap of/in us/ppo to/to read/vb into/in our/pp\$ authors/nns --/-- Nathaniel/np Hawthorne/np ./, for/in example/nn ./, and/cc Herman/np Melville/np --/-- protests/vbz against/in modernism/nn ./, material/jj progress/nn ./, and/cc science/nn which/wdt are/ber genuine/jj protests/nns of/in our/pp\$ own/jj but/cc may/md not/* have/hv been/ben theirs/pp\$\$./.

Faulkner's/np\$ total/nn works/nns today/nr ./, and/cc in/in fact/nn those/dts of/in his/pp\$ works/nns which/wdt existed/vbd in/in 1946/cd when/wrb Mr./np Cowley/np made/vbd his/pp\$ comment/nn ./, or/cc in/in 1939/cd ./, when/wrb Mr./np O'Donnell/np wrote/vbd his/pp\$ essay/nn ./, reveal/vb no/at such/jj simple/jj attitude/nn toward/in the/at South/nr-tl ./.

If/cs he/pps is/bez a/at traditionalist/nn ./, he/pps is/bez an/at eclectic/jj traditionalist/nn ./.

- (d) verwenden Sie jetzt den ganzen Text cg28. Stellen Sie sich vor Sie müssten einen automatischen Tagger (Wortartenerkennung) trainieren oder aber aus dem getaggten Text Regeln ableiten. Stellen Sie manuell einige Kontextmuster (Wörter vor oder nach dem Wort, das getaggt werden soll) fest die Ihnen brauchbar erscheinen. Etwa so: sie können in einem Lexikon jeweils die potentiellen Wortarten eines Tokens feststellen, im Englischen besteht dann u. a. sehr oft eine Ambiguität, dass dieselbe Wortform eine Lesart als Verb und eine als Nomen hat, z.B. **record, escort, insult**.

Welche Kontextmuster nützen zum Auflösen der Ambiguität.

- (e) verwenden Sie jetzt den Text cj2. Können Sie in diesem Text Phänomene feststellen, die in dem Artikel von Jack Halpern über die Segmentierung erwähnt wurden.

Aufgabe 2

- (a) Schauen Sie sich nochmals den Algorithmus findSplittings aus Aufgabe 5 an. Schreiben Sie in Ihrer eigenen Notation einen vereinfachten Algorithmus auf, der Komposita erkennt, die jeweils aus zwei einfachen Nomen ohne Fugenelement gebildet werden (Singular oder Pluralform). Beispiel Blumentopf. Gehen Sie davon aus, dass Sie das Lexikon als ungeordnete Liste von Nomen zur Verfügung haben.
- (b) Wenn Ihr Lexikon 100,000 Einträge hat Wie viele Stringvergleiche benötigen Sie dann

pro Durchlauf Ihres Algorithmus wenn ein durchschnittlicher String den Sie überprüfen 10 Buchstaben lang ist. Was könnten Sie tun um Stringvergleiche einzusparen.

Aufgabe 3

Lesen Sie den Text von Grefenstette und Tapanainen: *What is a word, what is a sentence* (Link auf der Kursseite).

Verwenden Sie das angegebene Verfahren wie folgt auf den Texten des Brown Korpus:

Versuchen Sie für jede der sechs Regeln auf Seite 7 des Artikels 3 Belegstellen im Brownkorpus zu finden. Geben Sie einen Zeilen-Dokument-Code für die Belegstellen an, so dass Sie diese wieder finden können.