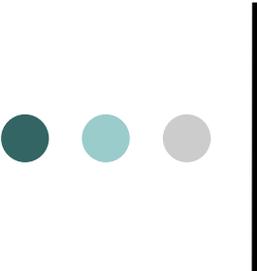


Part-of-Speech- Tagging

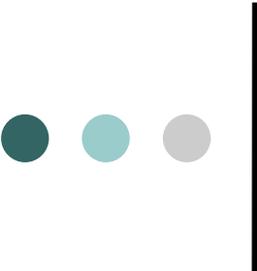
In: Einführung in die Computerlinguistik
Institut für Computerlinguistik
Heinrich-Heine-Universität Düsseldorf
WS 2004/05
Dozentin: Wiebke Petersen



Tagging...

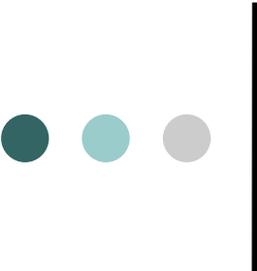
Was ist das?

- Tag (engl.): Etikett, Kennzeichen, Markierung
- Tagging: Das Zuweisen von Markierungen zu einzelnen Einheiten (Tokens)
- Part-of-Speech-Tagging (PoS-Tagging): Wortart-Annotierung



Verfahren... Regelbasiert

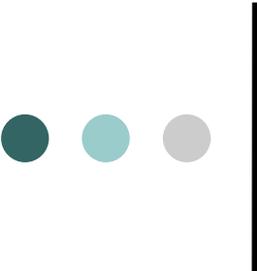
- Regelbasierte (transformationsbasierte, symbolische) Ansätze: Man stellt Regeln auf, die am Text abgearbeitet werden
- Vorteile:
 - Regeln lassen sich von Hand aufstellen
 - bereits wenige Regeln liefern Ergebnisse
- Nachteile:
 - Regeln sind Korpus- und Sprachspezifisch
 - Ein gutes System erfordert hohen Aufwand



Verfahren...

Statistisch

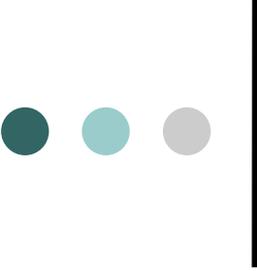
- Statistische (stochastische) Ansätze:
Das Programm lernt anhand
(annotierter) Trainingsdaten
 - Vorteil: Einfache Anpassung an neue Sprachen und Korpora durch erneutes Trainieren
 - Nachteil: Ein solches Programm ist schwer zu implementieren



Tokenisierung...

Wozu?

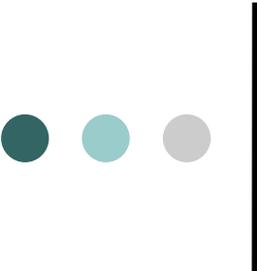
- Problem: Um einzelne Tokens zu annotieren, muss man sie haben
- Wortgrenzen: Leerzeichen, Satzzeichen
- Aber:
 - Im Englischen bestehen Begriffe oft aus mehreren „Wörtern“: „*soft ice*“
 - Satzzeichen markieren nicht unbedingt das Wortende: „*Heinrich-Heine-Universität*“
 - Sie können auch zum Wort gehören: „*bzw.*“



Tokenisierung...

Mögliche Probleme

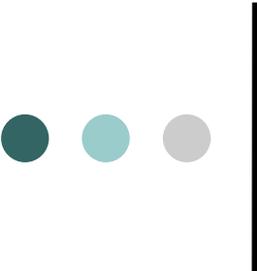
- Es gibt ein Problem:
 - „...*gut oder schlecht.*“: Punkt markiert das Satzende
 - „...*gut bzw. schlecht*“: Punkt gehört zum Wort
 - „...*gut, schlecht usw.*“: Punkt gehört zum Wort und markiert das Satzende



Tokenisierung...

Baselines

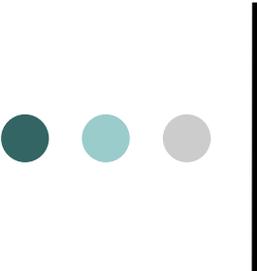
- Eine Baseline ist die Genauigkeit bei einer sehr einfachen Methode
- Bestimmung des Geschlechts: Alle als „*männlich*“ markieren – 50% richtig
- Bei Bindestrich an Zeilenende immer zusammenführen: 95% richtig
- Punkt bei nachfolgendem Leerzeichen und Großbuchstaben als Satzende markieren: 50-90% richtig



Tokenisierung...

Regelbeispiel (1)

- Auf Zeichensequenzen, die auf einen Punkt enden, werden folgende Regeln angewendet:
 - Folgt dieser Sequenz ein kleingeschriebener Buchstabe, ein Komma oder ein Semikolon, ist es eine Abkürzung
 - Ist das Wort eine kleingeschriebene Zeichenkette und existiert das gleiche Wort im Lexikon ohne Punkt, ist es keine Abkürzung



Tokenisierung...

Regelbeispiel (2)

- Beginnt die Sequenz mit einem Großbuchstaben, ist keine bekannte Abkürzung, findet sich im Korpus ohne abschließenden Punkt und besitzt nur eine sehr geringe Häufigkeit, ist es vermutlich ein Eigenname
- Ansonsten ist es eine Abkürzung
- Zusammen mit einer Liste häufig verwendeter Abkürzungen: Genauigkeit über 99%
- Realisierung z.B. mit regulären Ausdrücken

(Aus: Grefenstette, Tapanainen: *What is a Word, What is a Sentence? Problems of Tokenization*, 1994)

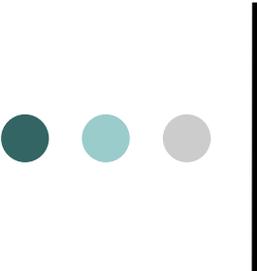


Tokenisierung...

Statistisches Beispiel

- Unsupervised learning – Lernen ohne Trainingsdaten
- Erster Durchlauf: Sammlung von statistischen Informationen, z.B. häufig mit Punkt vorkommende Wörter (Indikator für Abkürzungen), ohne Punkt vorkommende Wörter (Indikator gegen Abkürzungen), kleingeschriebene Wörter (nach Punkt groß geschrieben → Punkt markiert Satzende) usw.
- Zweiter Durchlauf: Berechnung der Wahrscheinlichkeiten anhand der gesammelten Daten

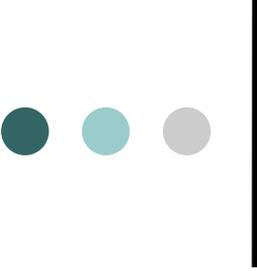
(Aus: Schmid: *Unsupervised Learning of Period Disambiguation for Tokenisation*, 2000)



Wortart-Tagging...

Mögliche Probleme

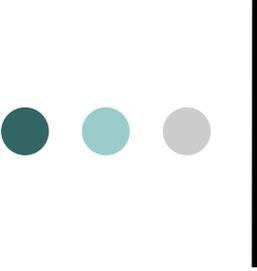
- Ambiguitäten:
 - *Wenn hinter Fliegen Fliegen fliegen, fliegen Fliegen Fliegen hinterher.*
 - *Ich entscheide es mit [meinen Kollegen].*
- Unbekannte Wörter:
 - *unwiderstehlich*
 - *Mark Twain ist ein US-Amerikanischer Schriftsteller.*
- Beides:
 - *Mein Name ist Kurz.*



Wortart-Tagging...

TnT

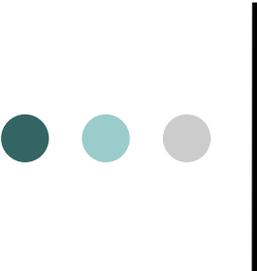
- Ein an der Universität Saarbrücken entwickelter Tagger
- TnT steht für Trigrams'n'Tags
 - Trigramme: Für die Bestimmung der Wortklasse werden die beiden vorausgehenden Wörter herangezogen
 - Ebenso gibt es Unigramme, Bigramme usw.
- Der Tagger arbeitet statistisch
- Für das Deutsche ist der Tagger auf einem 350.000 Tokens großen Korpus trainiert (Negra-Korpus)



Wortart-Tagging...

TreeTagger

- Ein an der Universität Stuttgart entwickelter Tagger
- TreeTagger, da mit Entscheidungsbäumen gearbeitet wird
 - Gibt es mehrere mögliche Entscheidungswege, so werden diese als Baum verarbeitet, an dessen Zweigen Wahrscheinlichkeiten notiert sind



Wortart-Tagging...

Morphy

- Ein an der Uni Paderborn entwickelter Tagger
- Lexikon mit 50.000 Stamm- und 350.000 Vollformen
- Morphologische Analyse
- (Statistische) Wortklassenprognose für unbekannte Wortformen