Non-lexicalized treebank grammars

- Wort Kategorien anstatt einzelner Wörter
  - weniger Informationen
  - leichter zu bauen

Charniak (1996)

- Verwendet:
  - Penn Treebank
  - part of speech, phrasal categories
  - maximum likelihood PCFG
- Verwendet nicht:
  - smoothing, rule collapsing
  - ungesehene sätze (ignoriert sie)
- Ergebnis erstaunlich gut

Pereira and Schabes (1992)

- Teilweise unüberwachtes lernen
- CNF, 15 Nichtterminale, 45 part of speech tags als Terminale
- Training auf Treebank Sätzen
  - Nichtterminale werden ignoriert
  - Treebank bracketing
- Input:
  - unbracketed: 37% der Test Sätze korrekt
  - bracketed: 90% der Test Sätze korrekt

Data-Oriented Parsing

- Alternative zu Grammatik basierten Modellen
- Statistik direkt über Teile eines Baums
  - Beispiel 12.28 und 12.29 auf Seite 446
- Parsing: Monte Carlo Simulation
  - Wahrscheinlichkeit eines Vorkommens wird aus zufälligen Beispielen geschätzt
  - Kann beliebig genau werden mit großer Anzahl an Beispielen, wird jedoch langsam

Data-Oriented Parsing

- Ähnlich zu MBL
  - Vorhersagen direkt durch den Korpus, aber DOP über den kompletten Korpus
- Unterschied zu PCFGs:
  - Baumteile können größer sein
  - Probabilistic Tree Substitution Grammar (PTSG)

Lexikalische Modelle

- Herleitungsgeschichte
  - History Based Grammars (HBGs)
  - Spatter

History Based Grammars

- Erforscht durch Experimente von IBM
  - Black et al. 1993
- Nutzen:
  - Ableitungsbaum
  - lexikalische und andere Informationen
- Annahme das alle vorherigen Entscheidungen die aktuelle beeinflussen
  - Entscheidungsbäume für den Herleitungsbaum
- Eigener treebank
  - Nutzt nur die 3000 häufigsten Wörter

Spatter

- Startet mit Wörtern und bildet die Struktur über sie
- Entscheidungsbäume wie bei black et al
- Parse Baum:
  - words, tags, labels und extensions
  - right ist das linkeste Kind, left das rechteste
  - up alle Kinder dazwischen
  - unary nur ein Kind
  - root

Spatter

- Modell benutzt folgende Fragen:
- Was ist X an dem (aktuellem Knoten/Knoten (½) zu dem (linken/rechten))?
- Was ist X an dem aktuellem Knoten (erste/zweite) (linkeste/rechteste) Kind?
- Wieviele Kinder hat der Knoten?
- Was ist der Bereich des Knoten in Worten?
- [Für Tags:] Was sind die 2 vorherigen POS tags?

Dependency-based models

- Collins (1996; 1997)
- lexikalische Abhängigkeiten
  - Wörter (B), Abhängigkeiten (D)
  - P(t|s) = P(B,D|s= P(B|s) x P(D|s,B)
  - Tagging unabhängiger Prozess
- Abstand zwischen Wörtern