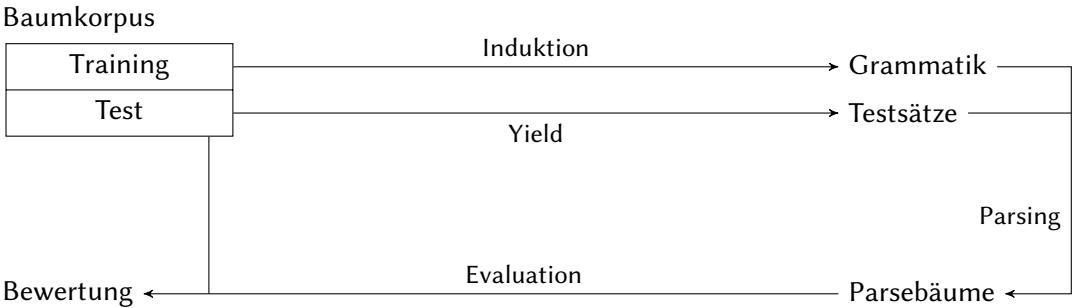
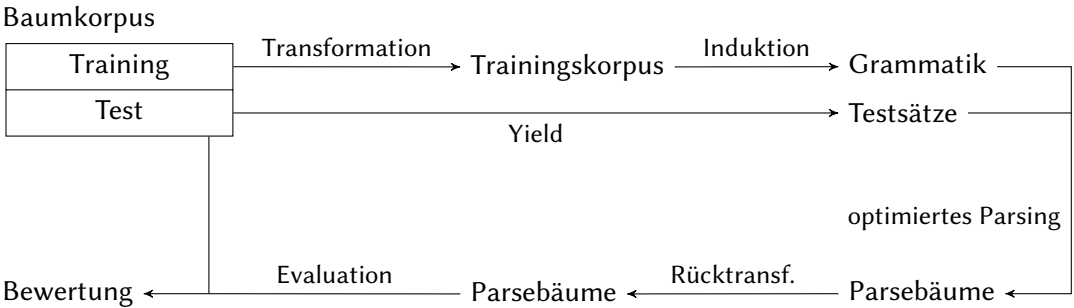
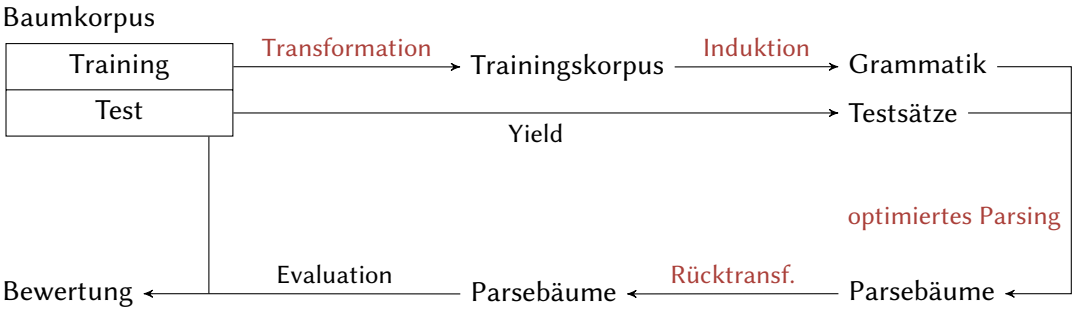
Zielstellung



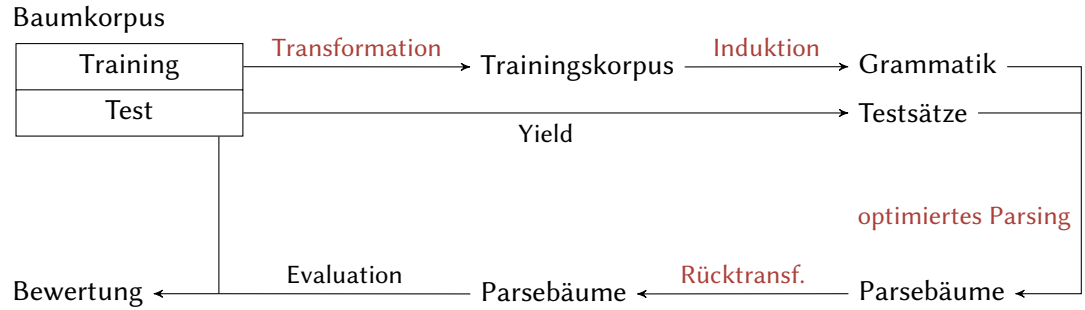
Zielstellung



Zielstellung



Zielstellung



Kriterien

- Genauigkeit/F1-Measure
- Laufzeit

Aufgabenstellung

Details: siehe Ablaufplan und Vorstellungen der einzelnen Teilaufgaben
Pflichtaufgaben:

1. Grammatikinduktion
2. Parser (bester Parsebaum) – Bottom-up (CYK) *oder* deduktiv (Knuths Algorithmus)
3. Korpustransformationen: triviales Unking und Debinarisierung

Wahlpflichtaufgaben (min. 3):

3. Korpustransformationen
   - Binarisierung und Markovisierung
   - Smoothing der induzierten Grammatik
4. Parseroptimierung
   - Pruning
   - n-best-Parsing
   - Heuristische Suche (A*)

Kolloquium/mündliche Prüfung (je nach Modul)

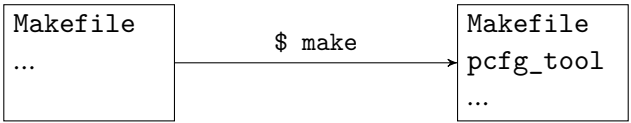
Auswertung

Code-Repository

```
Makefile
...
```

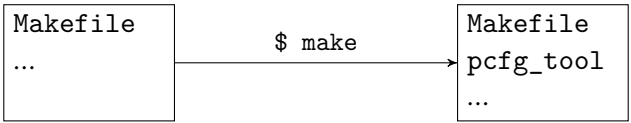
Auswertung

Code-Repository



Auswertung

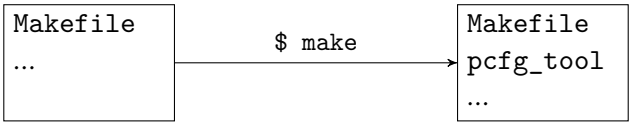
Code-Repository



- keine Dateien, die mit `.eval_` beginnen
- `pcfg_tool` ist ausführbar

Auswertung

Code-Repository



- keine Dateien, die mit `.eval_` beginnen
- `pcfg_tool` ist ausführbar

- `pcfg_tool` spricht Lösungen zu *allen* Aufgaben über Subkommandos an (wie git)
- Teilaufgabe nicht gelöst? → Exit-Code 22 bei Subkommando/Argument
- vollständige Kommandozeilenschnittstelle im Ablaufplan
- Überprüfung der Funktionalität durch Integration-Tests
- Automatisiertes Ranking im Wettbewerb

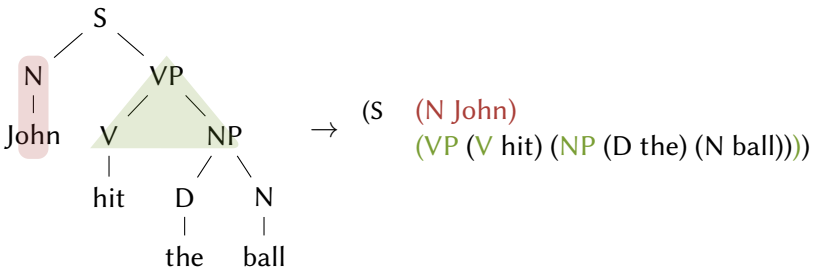
Formate

Generell: 1 Zeile = 1, Objekt → Streamverarbeitung

Formate

Generell: 1 Zeile = 1, Objekt → Streamverarbeitung

Konstituentenbäume: s-Expressions

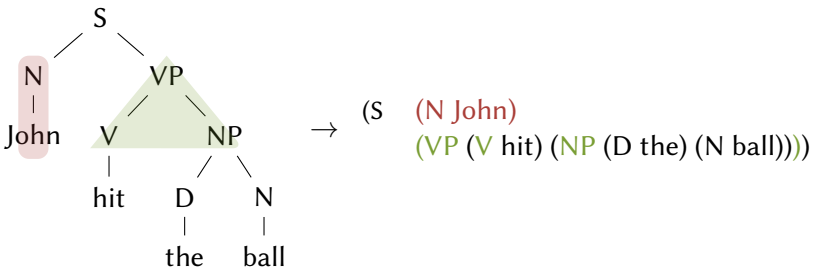


→ Nichtterminale und Terminale frei von Klammern (s. -LRB-, -RRB-)

Formate

Generell: 1 Zeile = 1, Objekt → Streamverarbeitung

Konstituentenbäume: s-Expressions



→ Nichtterminale und Terminale frei von Klammern (s. -LRB-, -RRB-)

Sätze: sind *tokenisiert*, i.e. “There’s no food”, he says. → “ There ’s no food ” , he says .

Material

Penn Treebank Wall Street Journal [Mar+94]

Trainingskorpus: Sections 00–18

- `training.mrg` – Konstituentenbäume für Grammatikinduktion

Grammatiken

- `grammar.rules`
- `grammar.lexicon`
- `grammar.words`

} aus binarisiertem Trainingskorpus induzierte Grammatik

Testkorpus: Sections 19–21

- `gold.mrg` – Konstituentenbäume für Evaluation
- `gold_b.mrg` – binarisierte Konstituentenbäume für Evaluation
- `testsentences` – Yield des Testkorpus für Evaluation

Organisation und Ablaufplan

Termin: Montag (zeitlich flexibel)

Tutorien (Teilnahme verbindlich) zu den folgenden Terminen:

- 19. 04. Einführungsveranstaltung, Vorstellung Aufgabe 1 (Grammatikinduktion)
- 10. 05. Auswertung Aufgabe 1, Vorstellung Aufgabe 2 (Parsing)
- 14. 06. Auswertung Aufgabe 2, Vorstellung Aufgabe 3 (Korpustransformationen)
- 21. 06. Vorstellung Aufgabe 4 (Parseroptimierung)
- 19. 07. Auswertung Aufgabe 3 und 4, Wettbewerb

Zu allen weiteren Terminen: *Möglichkeit* zur selbstständigen Arbeit

Abgabe der Aufgaben bis **Mittwoch, 23:59 Uhr MESZ** in der Woche vor der Auswertung

Kontakt

- Fragen
- Abgabe von Aufgaben
- Anmerkungen

richard.moerbitz@tu-dresden.de

Bis **10.05.2021**: Modul und Form der Prüfungsleistung

References I

[Mar+94] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz und B. Schasberger. „The Penn Treebank: Annotating Predicate Argument Structure“. HLT ’94. 1994.