Part Of Speech tags (POS)

- Klassifizierung der Wortart
- Typischerweise 10 bis 100 Klassen
- ADJ, ADV, V, N, PL, DET, ...
- Oft verwendet: „Brown corpus“ POS tags
- auch: STTS „Stuttgart-Tübingen TAG-Set“

POS–Markierung (‘tagging’)

- Voraussetzung: POS–Lexikon (ambig!)
- i.A. auf Stammformen basierend
- Es gibt einfache, gute Algorithmen
- ca. 96% Erkennungsrate
- naiv: ca. 80% Erkennungsrate
  (Englisch, Unigramm)
- Vorverarbeitungsschritt
- Desambiguierung beim Parsing

Statistische POS–Markierung

- Unigramm: Wahrscheinlichkeit eines POS-Tag für ein Wort

- n-Gramm: Wahrscheinlichkeit eines POS-Tag für ein Wort unter Kenntnis der n vorherigen POS (insb.: Bigramm und Trigramm)

- Problem der erforderlichen Datenmengen *sparse data problem*

Lexikalisierte TAGs, Baumfamilien

- Typischer Lexikoneintrag:
  ‘halten’ $\rightarrow$ *Verb–Transitiv*

- *Verb–Transitiv* ist ein **lexikalischer Typ**, der auf eine **Baumfamilie** abgebildet wird.

- Baumfamilien enthalten bis zu 40 Bäume

- Baumfamilien können als POS gesehen werden.

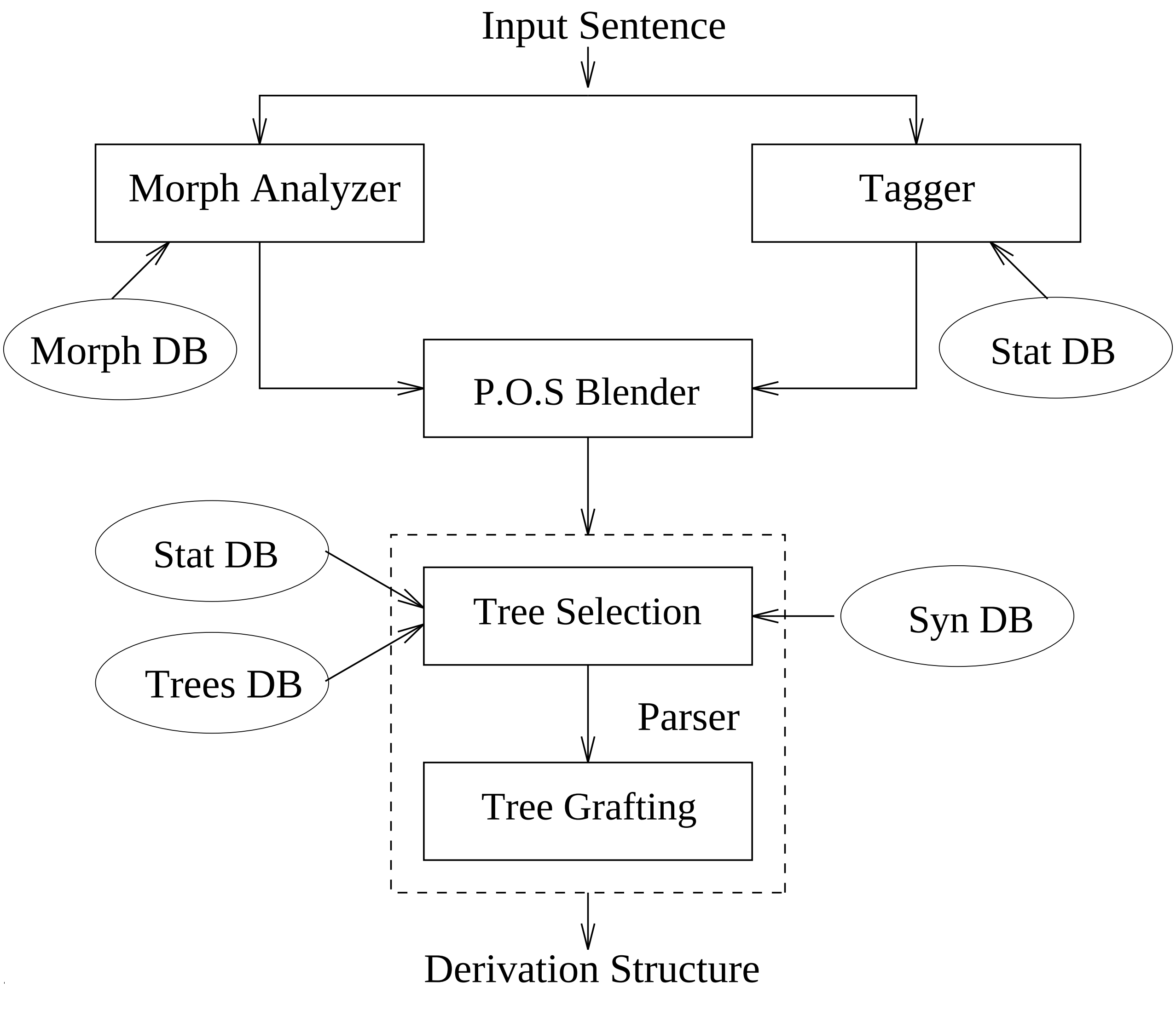
- In einer konkreten Ableitung ist jedem Wort ein spezifischer (elementarer) Baum zugeordnet.

- Elementare Bäume können als hochspezifische POS gesehen werden.

$\rightarrow$ **Supertags**, (Srinivas 1994)

XTAG

Die Parsingarchitektur des XTAG-Systems:



Daten für Supertags

- WSJ (Wall Street Journal):
  5000 Sätze, 21000 Worte

- Unigramm, Trigramm, Dependenzmodell

- Strukturelle Constraints:
  - Zahl der Substitutionsknoten
  - weitere Worte im Baum

- Problem der Datenabdeckung:
  daher nur POS/Supertag–Statistiken

Ergebnisse

- **Unigramm** Supertag-Modell:
- Erst Standard POS-Bestimmung
- Dann Parsingversuch mit den n besten (Unigramm) Supertags:
-

Top n Supertags	% Erfolg
$n = 1$	15%
$n = 2$	22%
$n = 3$	52%

- Beschleunigung: 87% bei Erfolg

Ergebnisse II:

- **Trigramm** Supertag-Modell
- Erst recht nur mit POS/Supertag-Paaren
- Bessere Werte für Einzelworte:
  68% korrekte Supertags
- Keine Angabe über korrekte Sätze.

Dependenz-basiertes Modell

- Auch hier nur POS/Supertag-Paare
- Darüber hinaus auch noch:
  - Liste von $+, -$; repräsentiert die Richtung abhängiger Supertags
  - Abhängige Supertags mit
    Distanz und
    Wahrscheinlichkeit
- Beispiel:
  $(V, \alpha_2), (-, +), \alpha_8, -1, 0.300$
  $(V, \alpha_2), (-, +), \alpha_8, +1, 0.374$
- α_8 *tritt mit Wahrscheinlichkeit 0.374 als direkter linker* $(+1)$ *Nachbar auf.*

Beispiel

Satz:

John	saw	a	man	through	the	telescope

Initiale Auswahl:

α_1	α_2	β_2	α_4	β_1	β_4	β_5
α_7	α_8	α_3	α_9	β_3	α_5	α_6
$\alpha_1 0$	$\alpha_1 1$	$\alpha_1 2$	$\alpha_1 3$	β_6	$\alpha_1 4$	$\alpha_1 5$

Ergebnis:

$\alpha_1 0$	α_2	α_3	α_4	β_6	α_5	α_6

- Jeder Kopf muß alle Dependenten finden.
- Jeder Dependent muß an einen Kopf gebunden sein.
- Keine zwei Abhängigkeits-Links dürfen sich kreuzen.

Dependenz-basiertes Modell

- Der Algorithmus konstruiert den wahrscheinlichsten Ableitungsbaum

- I.A. komplexes Problem ($O(n^3)$??), hier werden Heuristiken verwandt:
  greedy Algorithmus zur Maximierung der *akkumulierten Pfadwahrscheinlichkeit*

- Ergebnisse III:

	Anzahl	davon korrekt	% korrekt
Dependenzen	815	620	76.07%
Supertags	915	707	77.26%

- *almost parsing*

Supertags

- Supertags sind allgemein anwendbar für (vollständig) lexikalisierte Grammatikformalismen

- Beispiel: CCG

- Einschränkung:

  - bei LTAGs folgt aus Dependenz (fast) automatisch die Kombinationsoperation (Substitution bzw. Adjunktion)

  - bei CCGs muß die Kombinationsoperation (i.A.) separat bestimmt werden

Supertags vs. POS-Tags

- Supertags sind wesentlich feiner als Part-Of-Speech Tags
- Daher schlechtere ErkennungsrateP:
  15% vs. 80% (Unigramm)
  77% vs. 96% (bestes Verfahren)
- Aber mehr Information für Parsing
- Supertags können als Vorfilter im Parsing eingesetzt werden (schnelle Anwendung)

Example Based Learning

- EBL lernt „wahrscheinliche“ Ableitungsbäume

- Ein *Index* erlaubt Abstraktion vom konkreten Satz.

- Bei kontextfreien Grammatiken:
  abstrahiert von Ableitungsbäumen
  (= Phrasenstrukturbäumen)

- Bei TAGs:
  Ableitungsbäume enthalten Supertag-Zuweisung!

- Auch hier POS- und keine Wortstatistiken

EBL Beispiel

- Satz mit POS-Markierung:
  Show/V me/N the/D flights/N from/P Boston/N to/P Philadelphia/N.

- Indexbildung über POS:
  V N D N (P N)*

- Index ist regulärer Ausdruck,
  über Adjunktion ist Rekursion
  (im regulären Ausdruck als Kleene-Stern)
  unmittelbar erkennbar.

- Anpassen gelernter Beispielparses bei passendem Index (regulärem Ausdruck) erfordert gewissen Aufwand (*stapler*)

EBL Ergebnisse

- ATIS Korpus, 365 Trainings-, 100 Testsätze

System	Abdeckung	Ø Zeit (Sek.)
XTAG	100%	125.18
EBL (nur Zugriff)	80%	1.78
EBL+Stapler	70%	8.00
EBL+XTAG Parser	90%	62.93