Syntax-basierte maschinelle Übersetzung mit Baumübersetzern

Andreas Maletti

Leipzig — 28. April 2015

Maschinelle Übersetzung

Original

Übersetzung (GOOGLE TRANSLATE)

- The addressees of this paper are students and students will be in the audience are.

Maschinelle Übersetzung

Original

- Die Adressaten dieses Vortrags sind Studierende und im Publikum werden sich Studierende befinden.

Übersetzung (GOOGLE TRANSLATE)

- The addressees of this paper are students and students will be in the audience are.

Maschinelle Übersetzung

Original

- Die Adressaten dieses Vortrags sind Studierende und im Publikum werden sich Studierende befinden.

Übersetzung (GOOGLE TRANSLATE)

- The addressees of this paper are students and students will be in the audience are.
- To scientific lecture, a public discussion follows on.

Maschinelle Übersetzung

Original

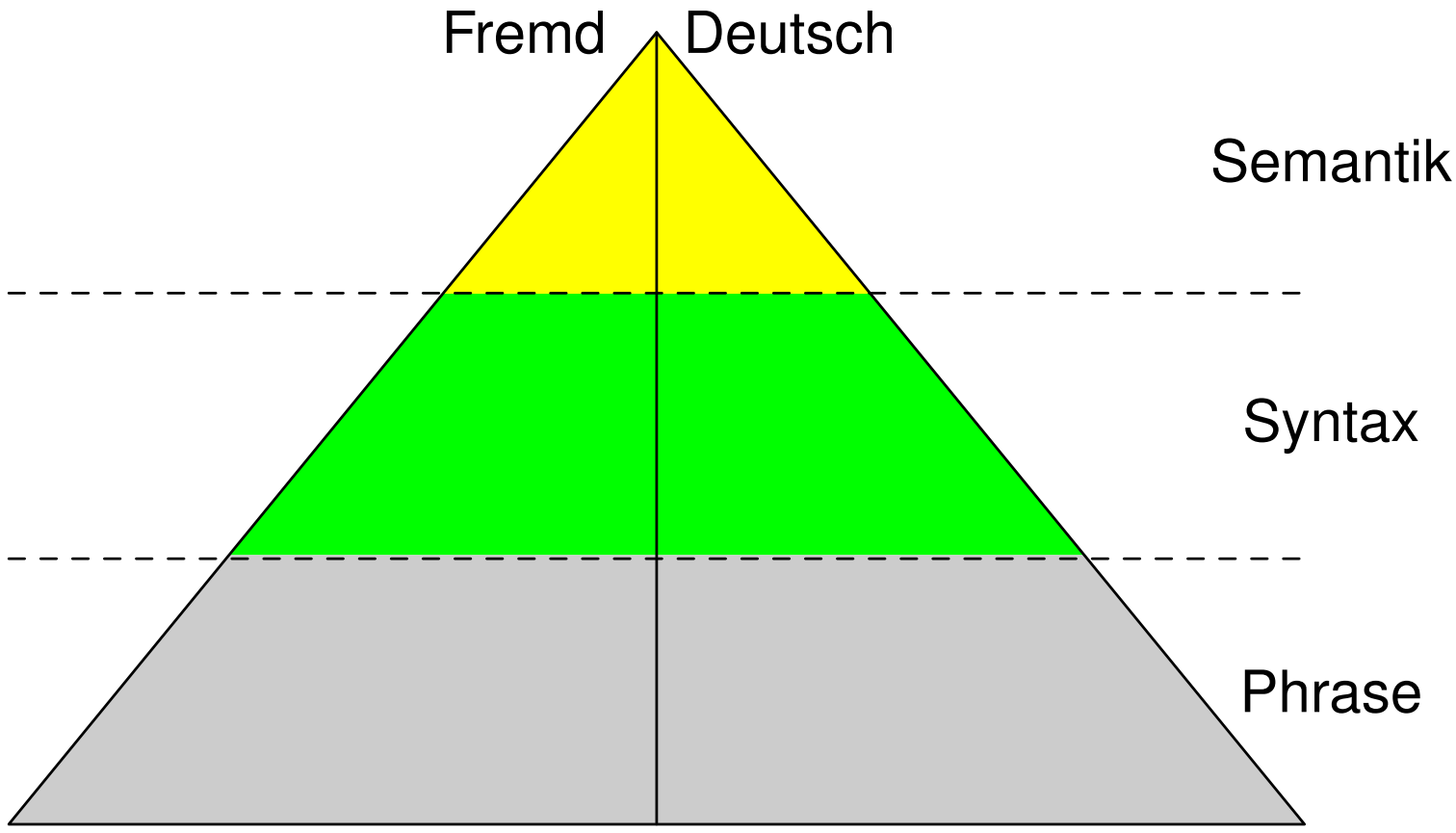
- Die Adressaten dieses Vortrags sind Studierende und im Publikum werden sich Studierende befinden.
- An den wissenschaftlichen Vortrag schließt sich eine öffentliche Diskussion an.

Übersetzung (GOOGLE TRANSLATE)

- The addressees of this paper are students and students will be in the audience are.
- To scientific lecture, a public discussion follows on.

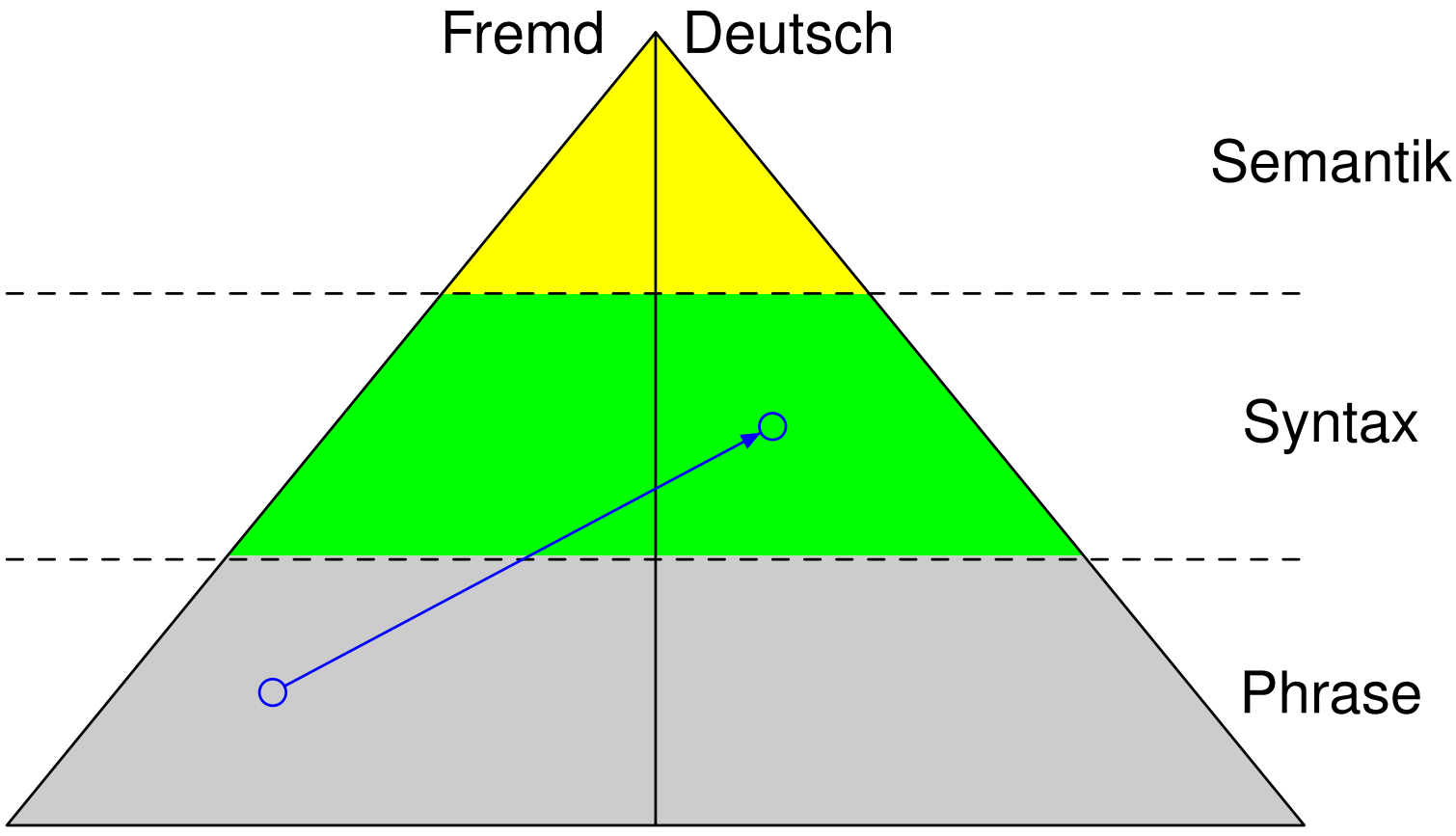
Maschinelle Übersetzung

Dreieck von VAUQUOIS:



Maschinelle Übersetzung

Dreieck von VAUQUOIS:



Übersetzungsmodell: “string-to-tree”

Maschinelle Übersetzung

Trainingsdaten

- paralleler Korpus
- Wortbeziehungen
- Syntaxbäume der deutschen Sätze (Zielsprache)

Maschinelle Übersetzung

Trainingsdaten

- paralleler Korpus
- Wortbeziehungen
- Syntaxbäume der deutschen Sätze (Zielsprache)

Paralleler Korpus

Linguistische Ressource mit Beispielübersetzungen (auf Satzebene)

Maschinelle Übersetzung

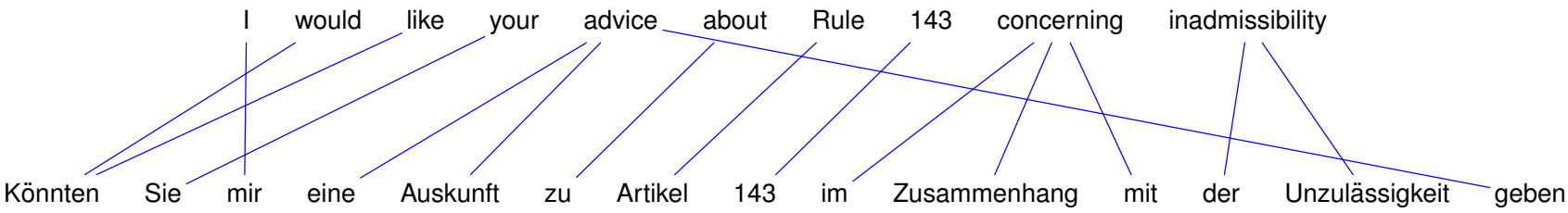
paralleler Korpus, Wortbeziehungen, Syntaxbaum

I would like your advice about Rule 143 concerning inadmissibility

Könnten Sie mir eine Auskunft zu Artikel 143 im Zusammenhang mit der Unzulässigkeit geben

Maschinelle Übersetzung

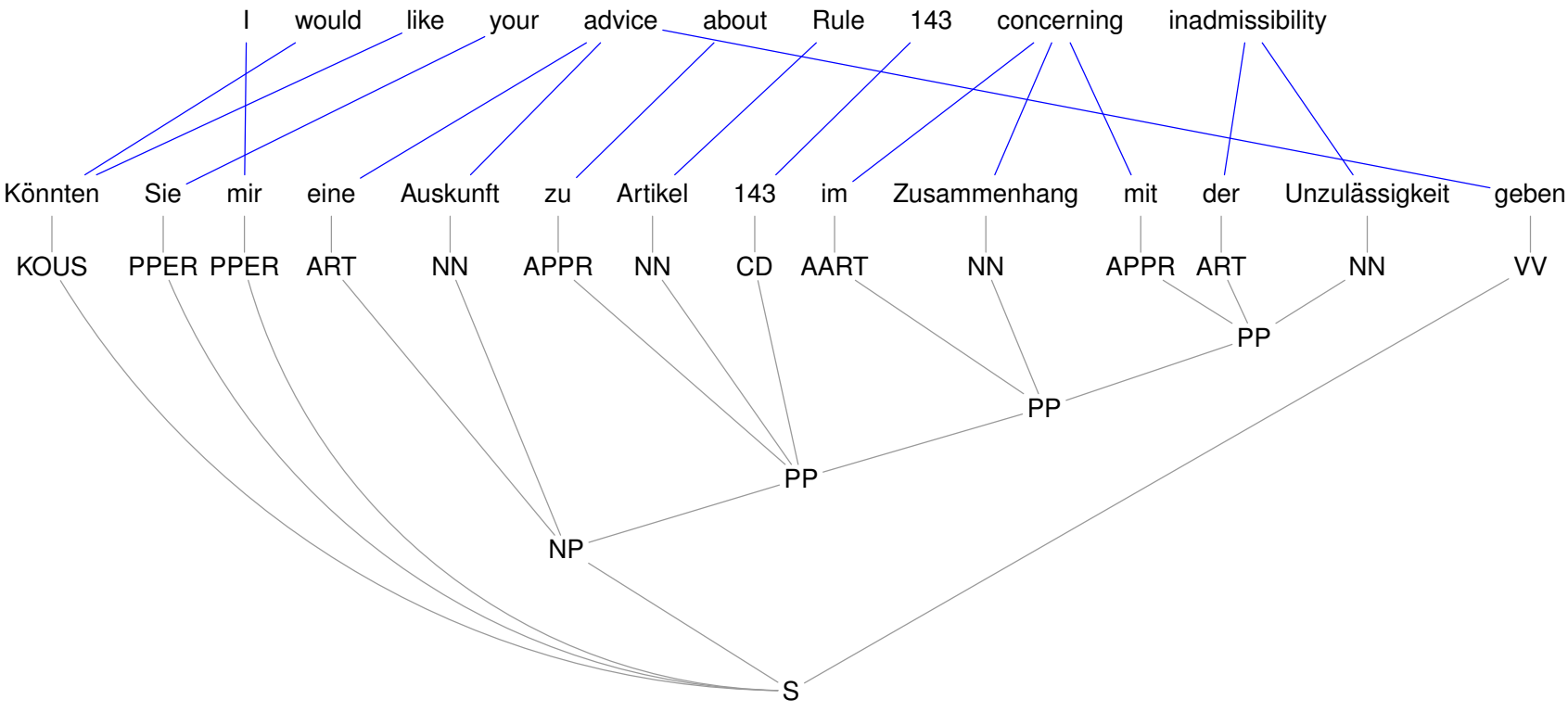
paralleler Korpus, Wortbeziehungen, Syntaxbaum



per GIZA++ [Och, Ney, 2003]

Maschinelle Übersetzung

paralleler Korpus, Wortbeziehungen, Syntaxbaum



per BERKELEY-Parser [PETROV et al., 2006]

Baumübersetzer

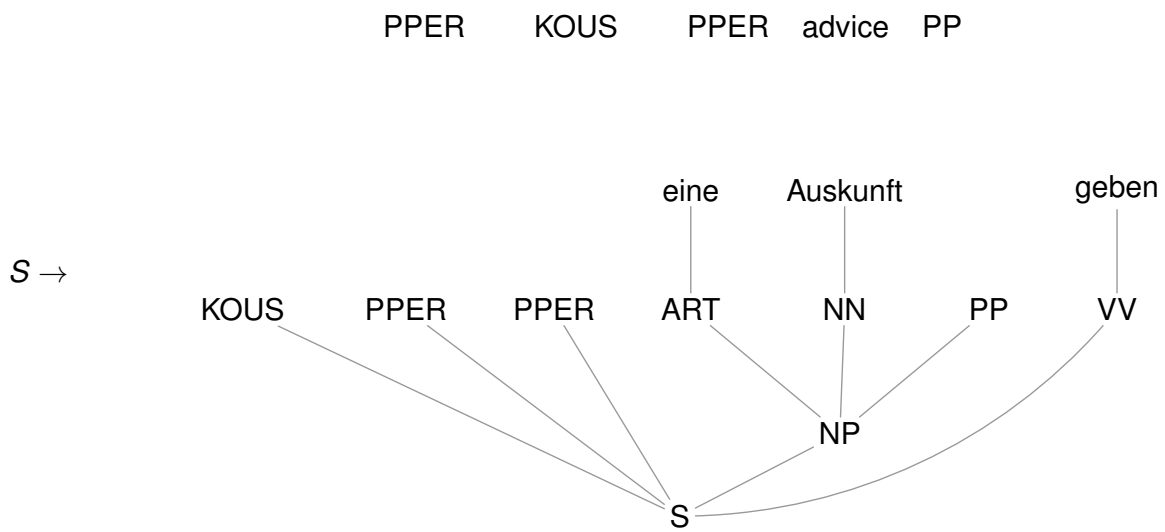
Erweiterter absteigender Baumübersetzer (STSG)

- Variante von [M., GRAEHL, HOPKINS, KNIGHT, 2009]
- Regeln der Gestalt NT $\rightarrow (r, r_1)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseite r_1 einer regulären Baumgrammatik

Baumübersetzer

Erweiterter absteigender Baumübersetzer (STSG)

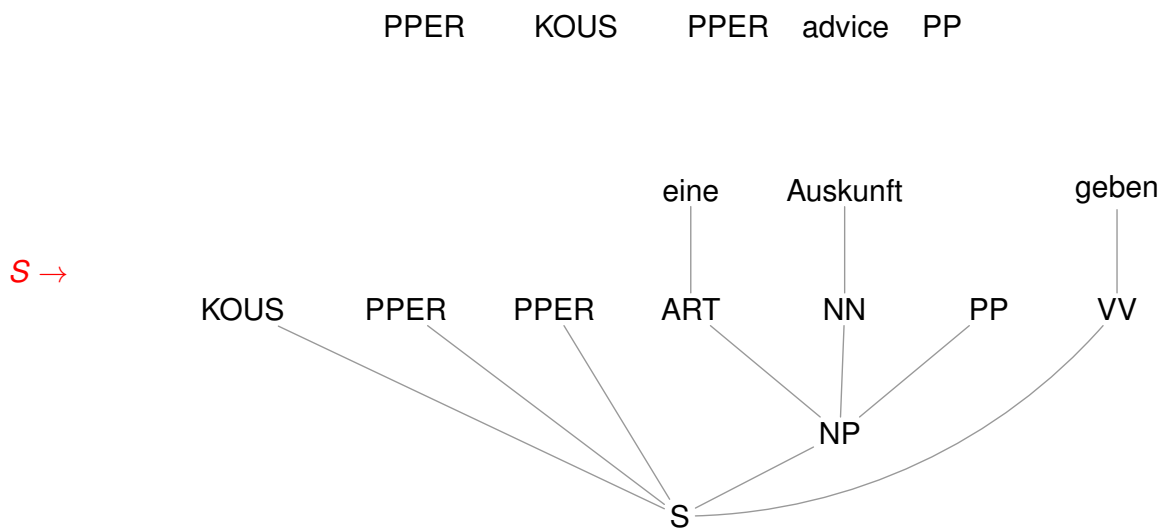
- Variante von [M., GRAEHL, HOPKINS, KNIGHT, 2009]
- Regeln der Gestalt NT $\rightarrow (r, r_1)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseite r_1 einer regulären Baumgrammatik



Baumübersetzer

Erweiterter absteigender Baumübersetzer (STSG)

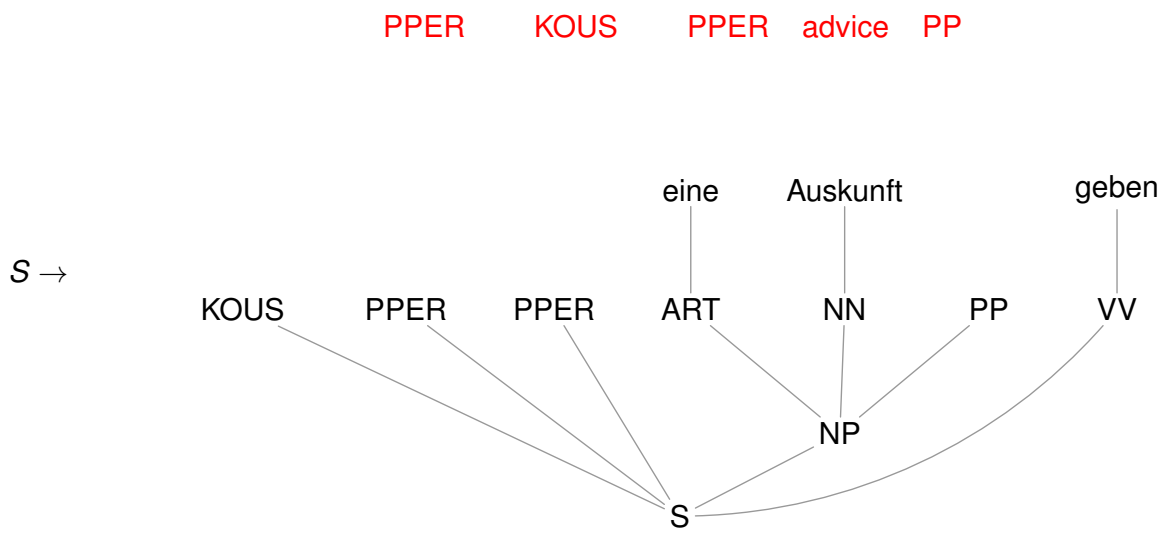
- Variante von [M., GRAEHL, HOPKINS, KNIGHT, 2009]
- Regeln der Gestalt NT $\rightarrow (r, r_1)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseite r_1 einer regulären Baumgrammatik



Baumübersetzer

Erweiterter absteigender Baumübersetzer (STSG)

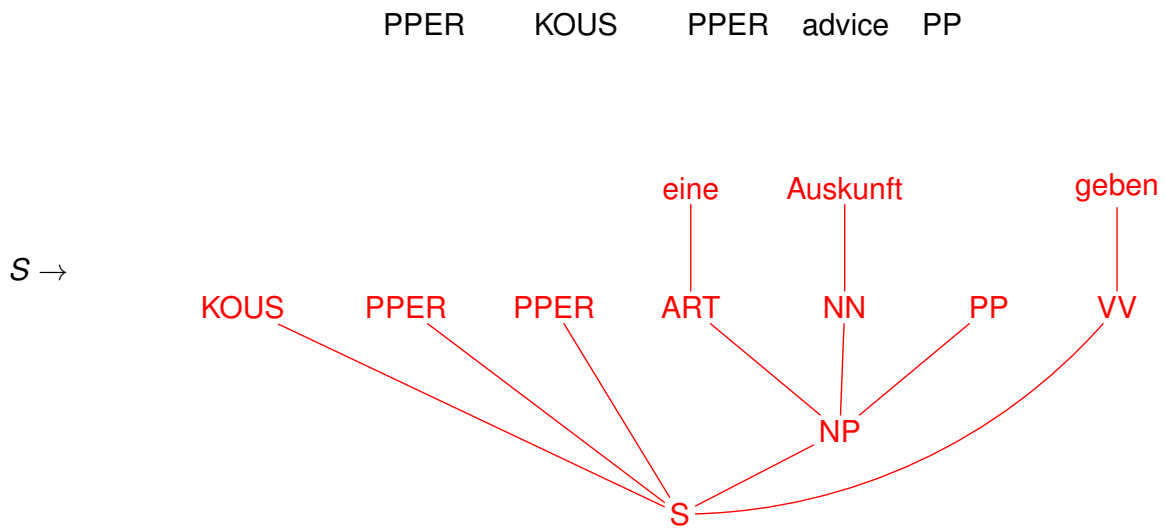
- Variante von [M., GRAEHL, HOPKINS, KNIGHT, 2009]
- Regeln der Gestalt NT $\rightarrow (r, r_1)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseite r_1 einer regulären Baumgrammatik



Baumübersetzer

Erweiterter absteigender Baumübersetzer (STSG)

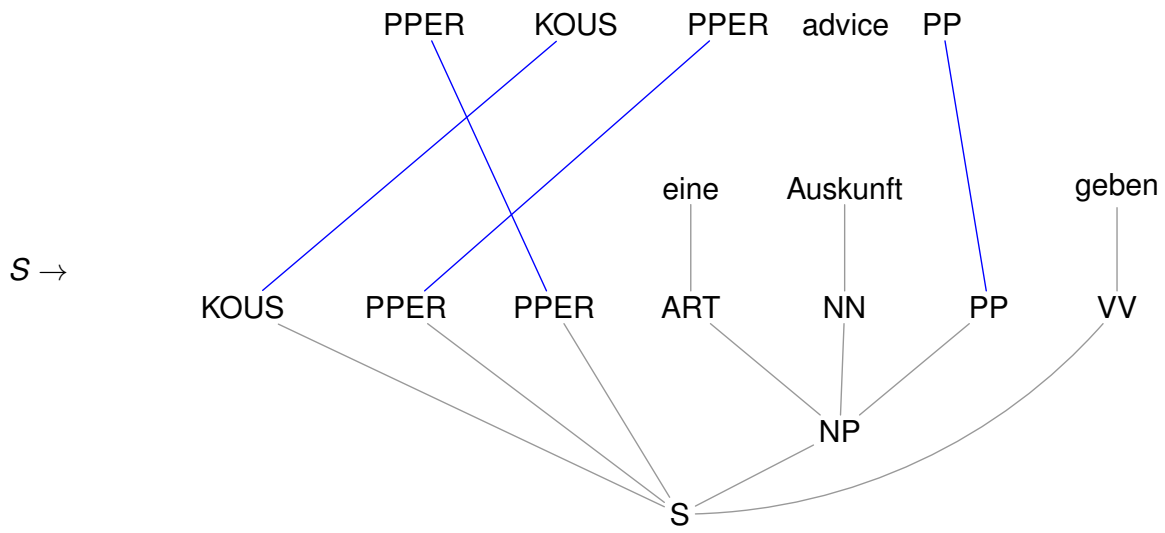
- Variante von [M., GRAEHL, HOPKINS, KNIGHT, 2009]
- Regeln der Gestalt NT $\rightarrow (r, r_1)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseite r_1 einer regulären Baumgrammatik



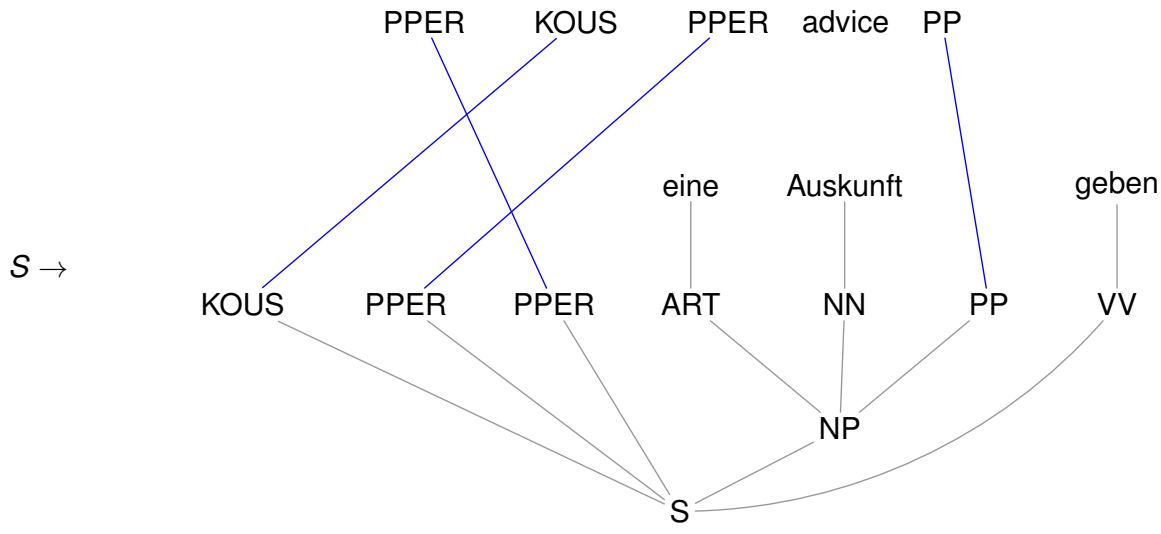
Baumübersetzer

Erweiterter absteigender Baumübersetzer (STSG)

- Variante von [M., GRAEHL, HOPKINS, KNIGHT, 2009]
- Regeln der Gestalt NT $\rightarrow (r, r_1)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseite r_1 einer regulären Baumgrammatik
- (bijektive) Synchronisation der Nichtterminale



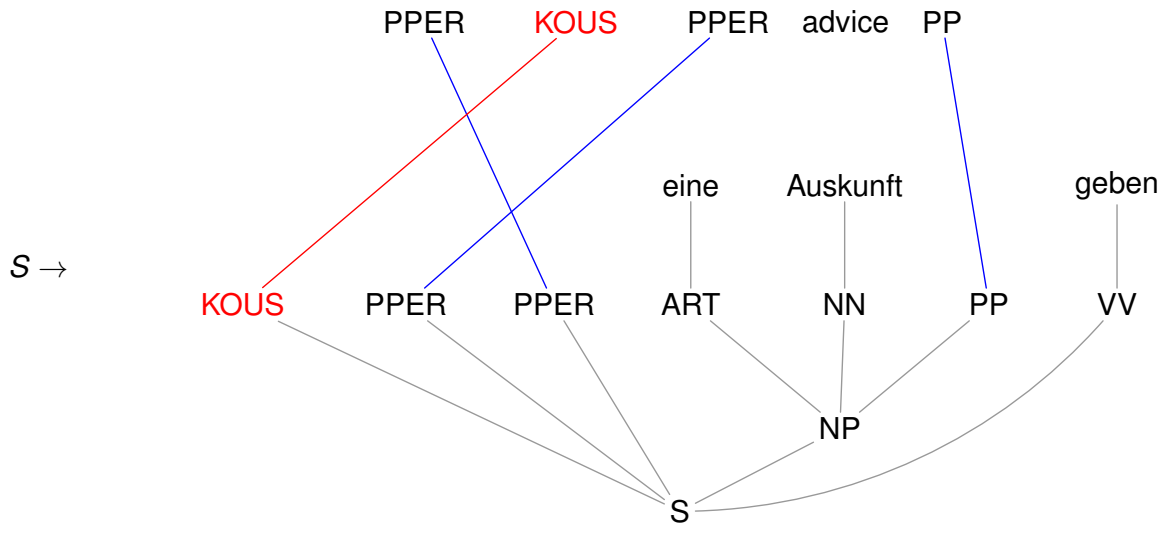
Baumübersetzer



Regelanwendung

1. Auswahl synchroner Nichtterminale

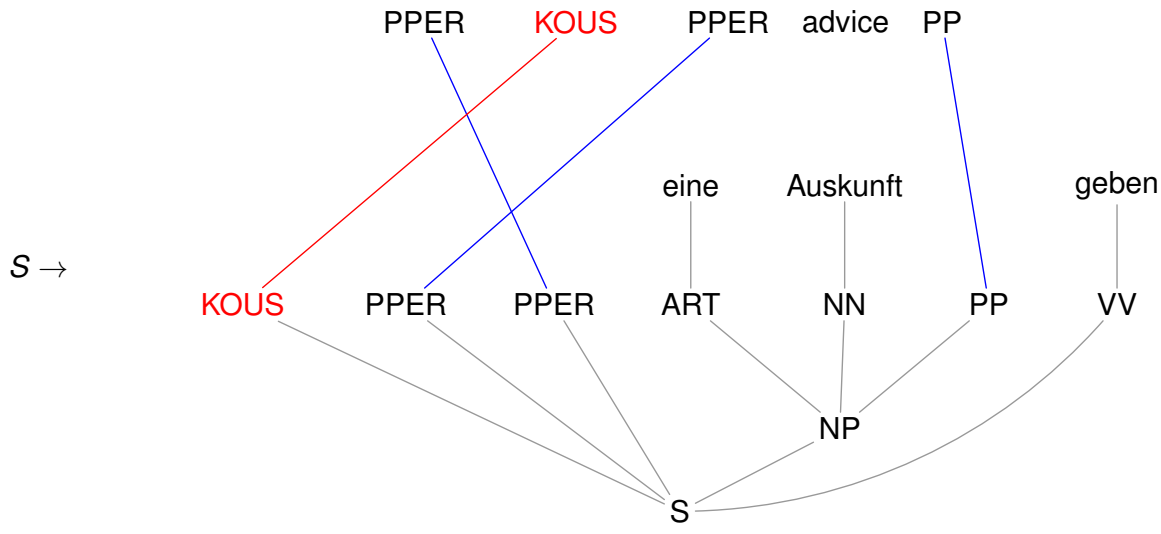
Baumübersetzer



Regelanwendung

1. Auswahl synchroner Nichtterminale

Baumübersetzer



Regelanwendung

1. Auswahl synchroner Nichtterminale
2. Auswahl einer passenden Regel

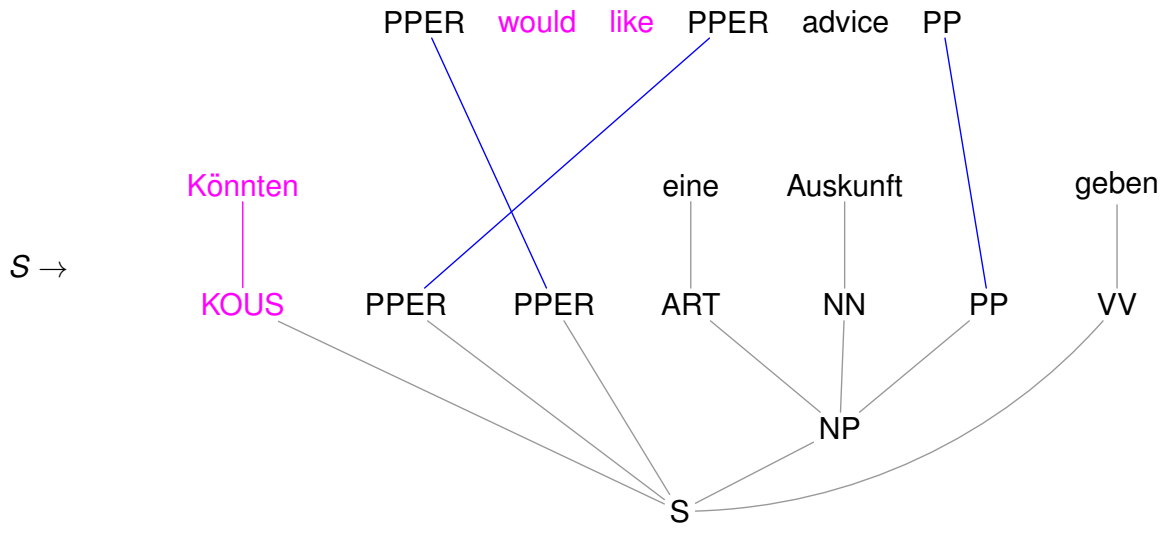
would like

KOUS →

Könnten

KOUS

Baumübersetzer



Regelanwendung

1. Auswahl synchroner Nichtterminale
2. Auswahl einer passenden Regel
3. Ersetzung auf beiden Seiten

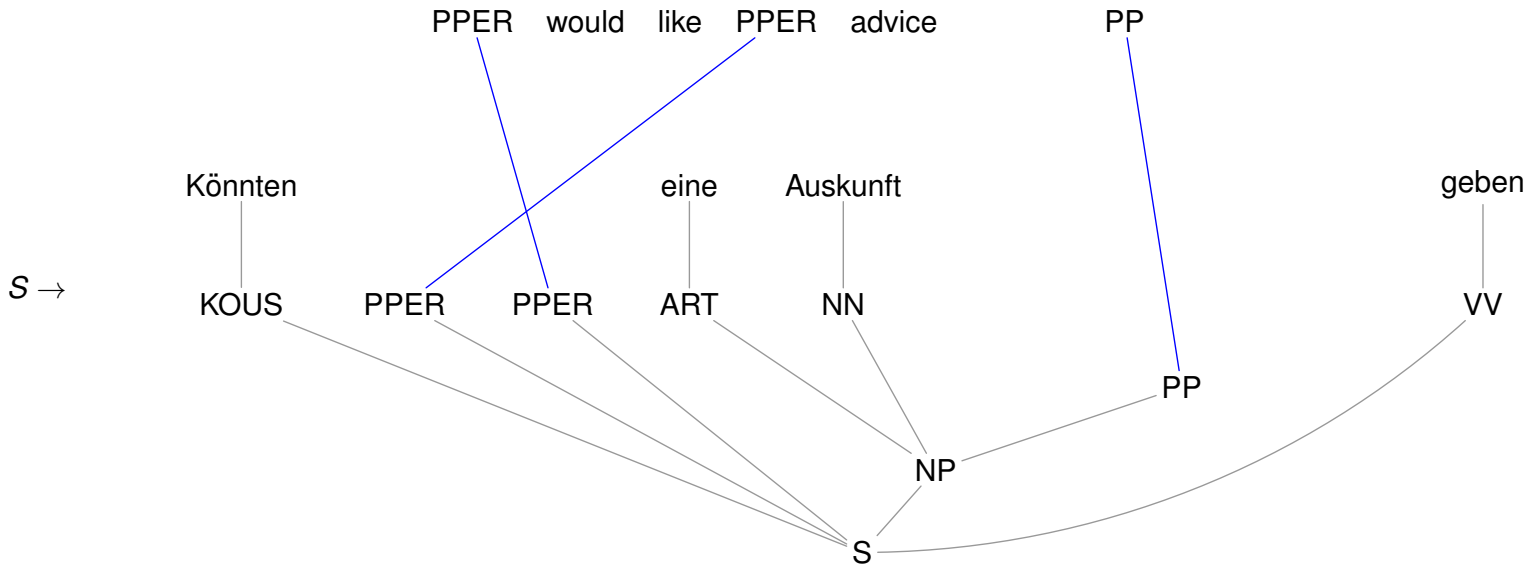
would like

KOUS →

Könnten

KOUS

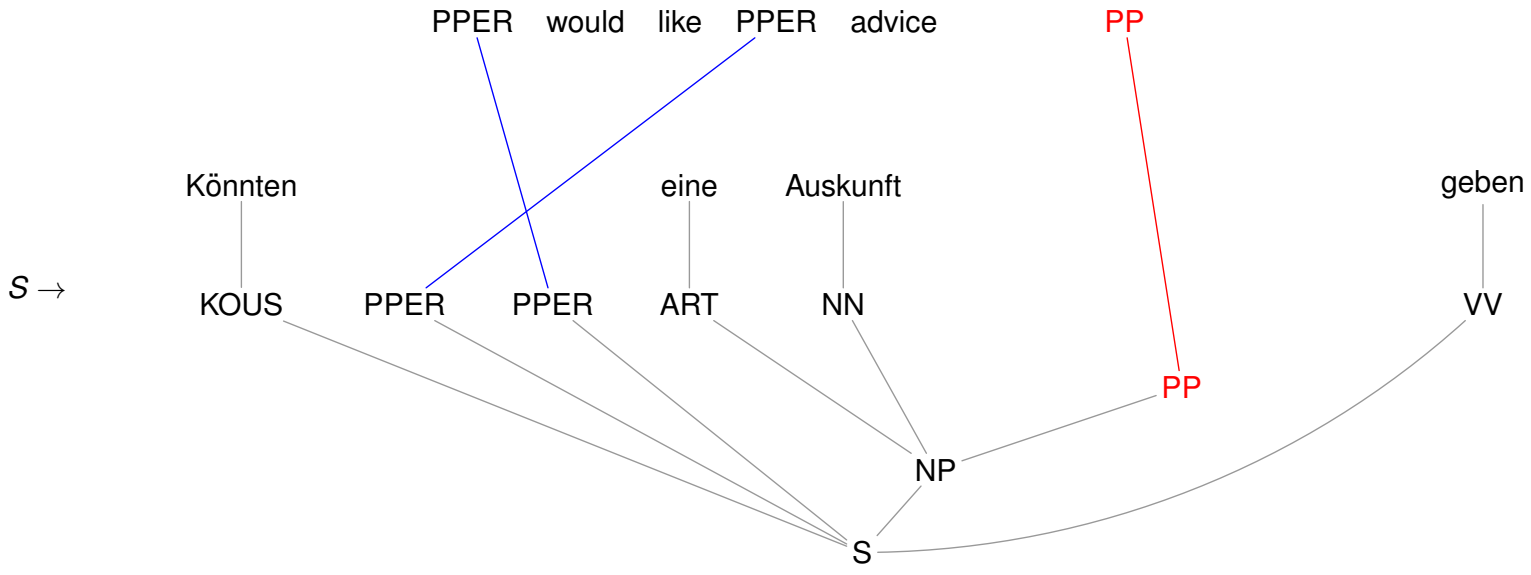
Baumübersetzer



Regelanwendung

1. Nichtterminal-Auswahl

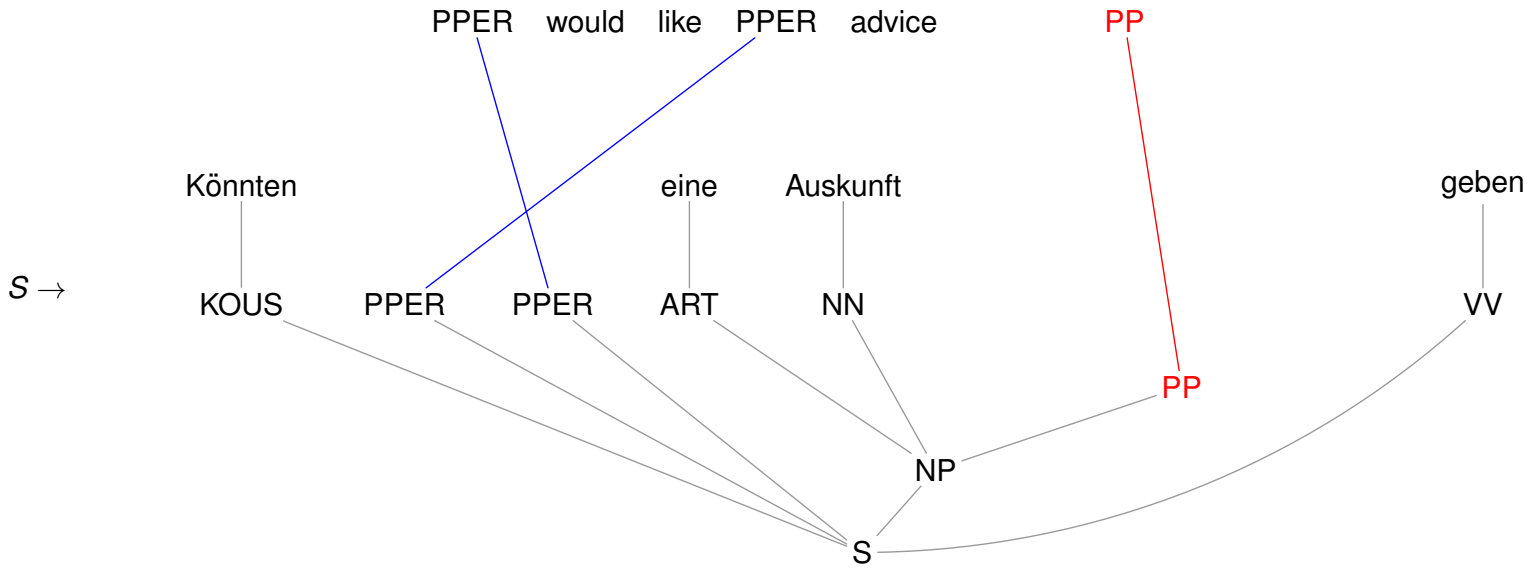
Baumübersetzer



Regelanwendung

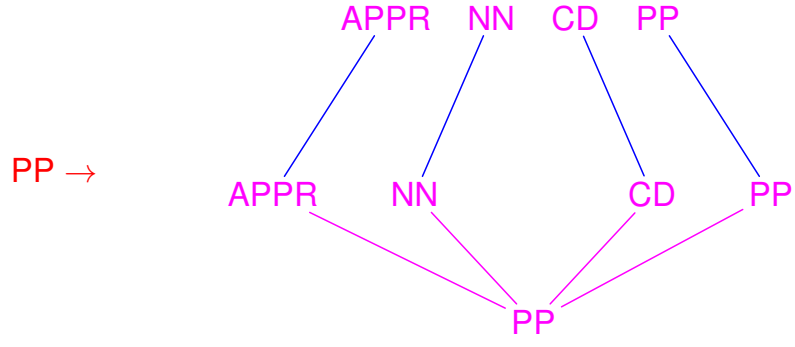
1. Nichtterminal-Auswahl

Baumübersetzer

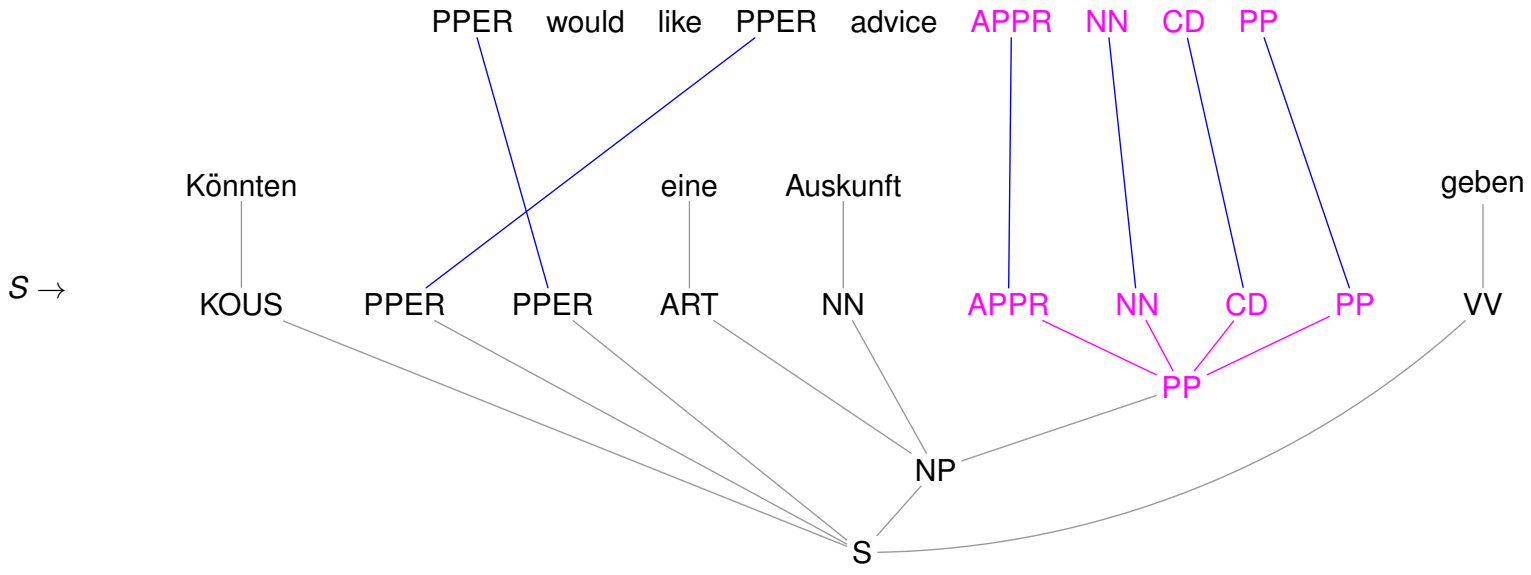


Regelanwendung

1. Nichtterminal-Auswahl
2. passende Regel

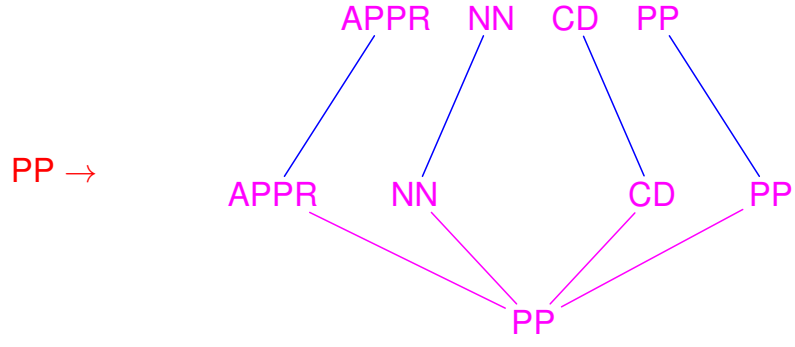


Baumübersetzer



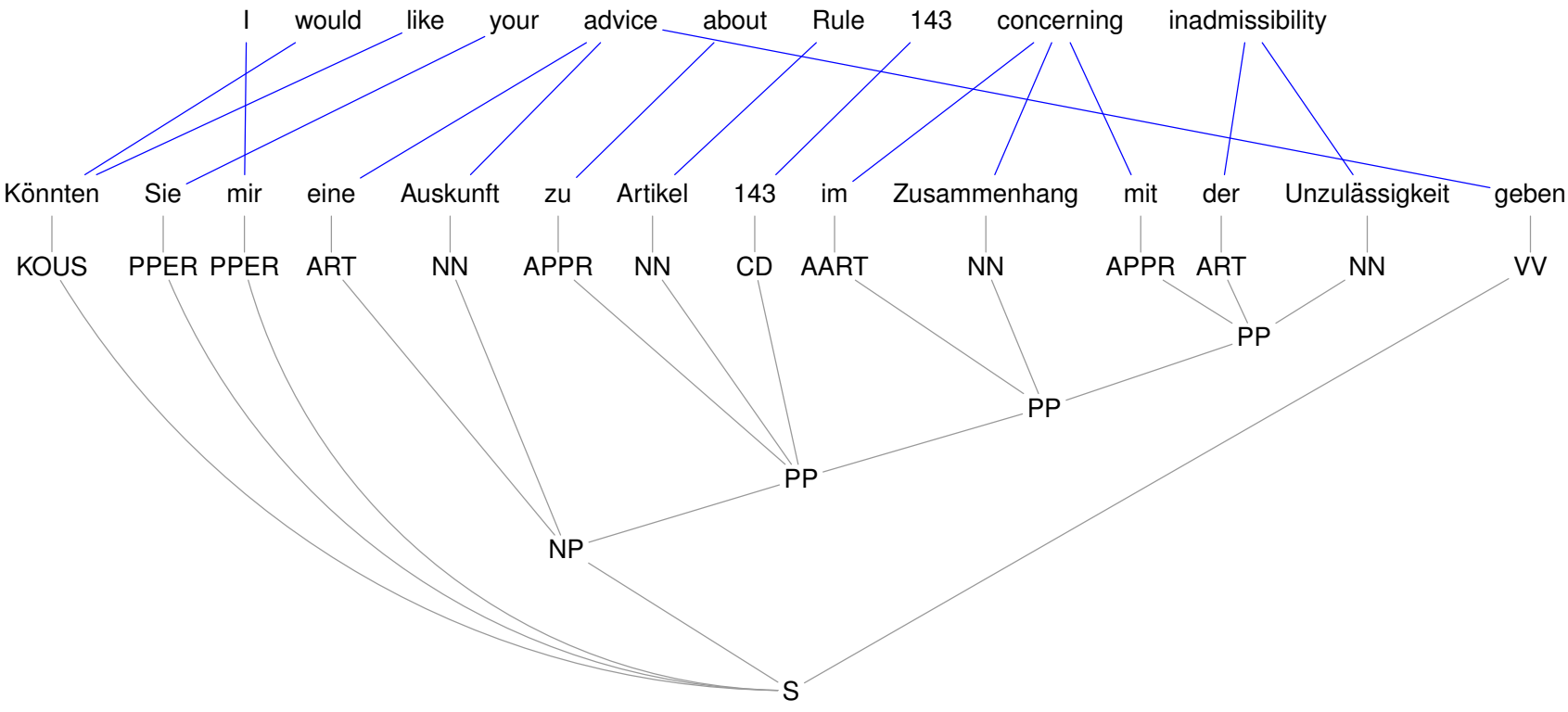
Regelanwendung

1. Nichtterminal-Auswahl
2. passende Regel
3. Ersetzung



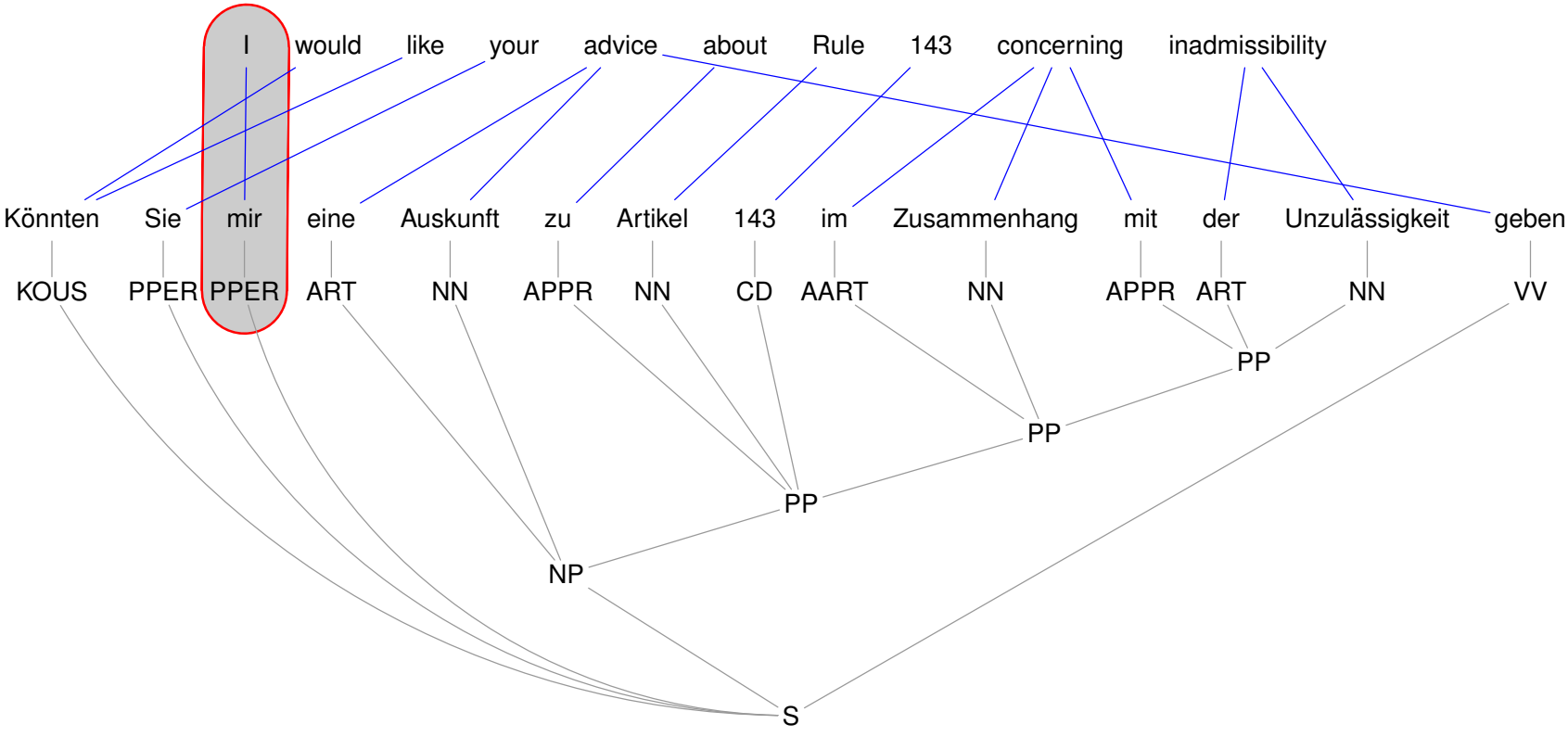
Regelextraktion

nach [GALLEY, HOPKINS, KNIGHT, MARCU, 2004]



Regelextraktion

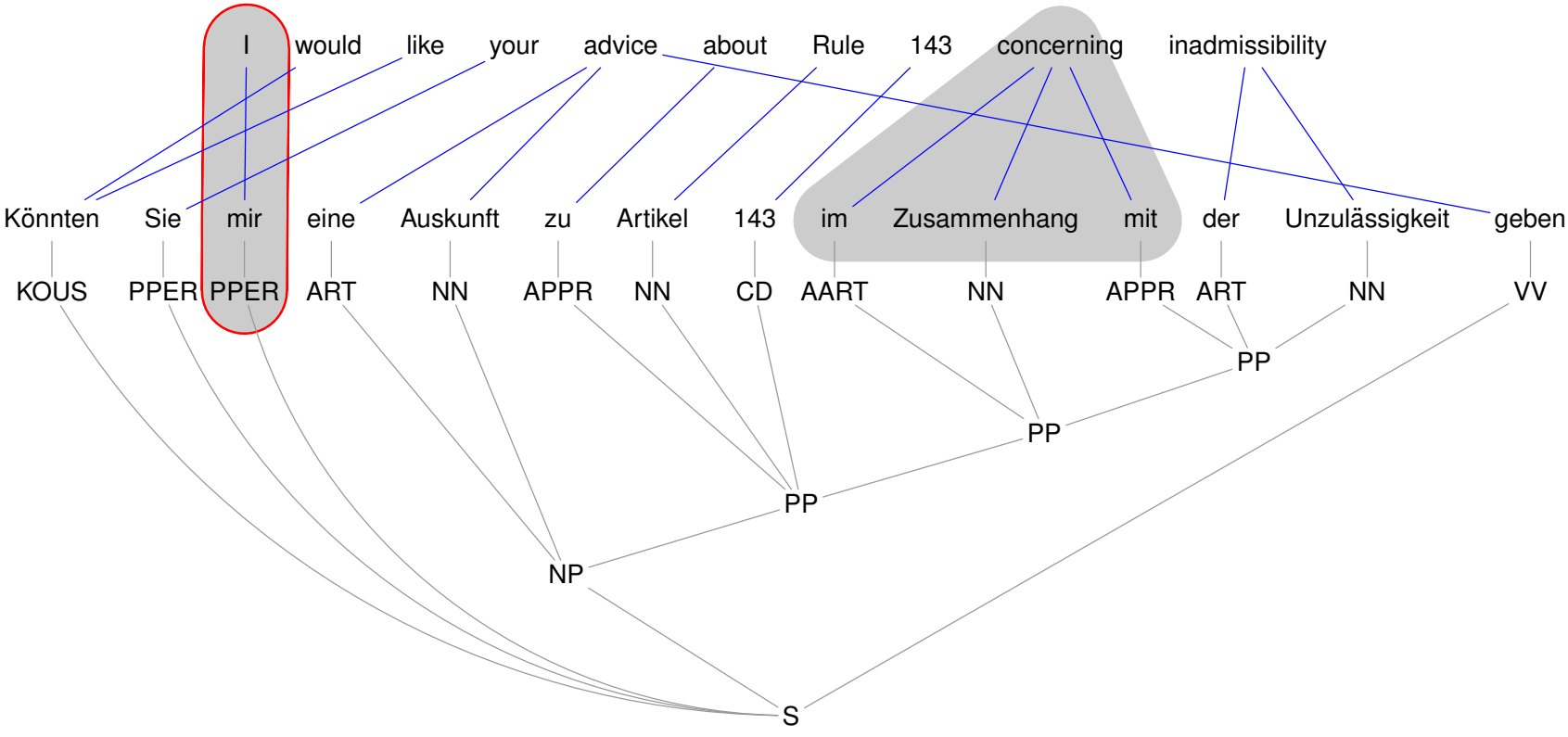
nach [GALLEY, HOPKINS, KNIGHT, MARCU, 2004]



extrahierbare Regeln rot umrandet

Regelextraktion

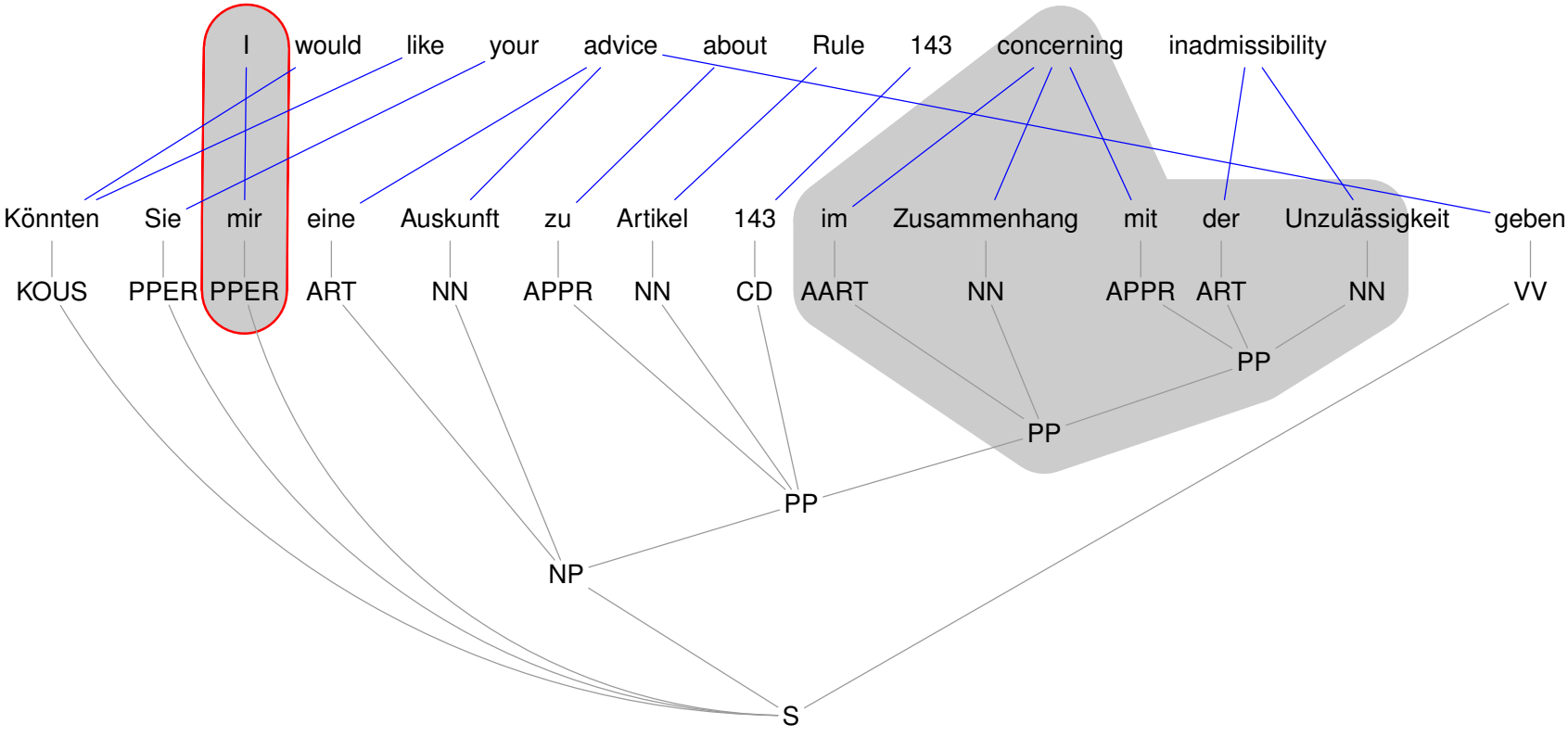
nach [GALLEY, HOPKINS, KNIGHT, MARCU, 2004]



extrahierbare Regeln rot umrandet

Regelextraktion

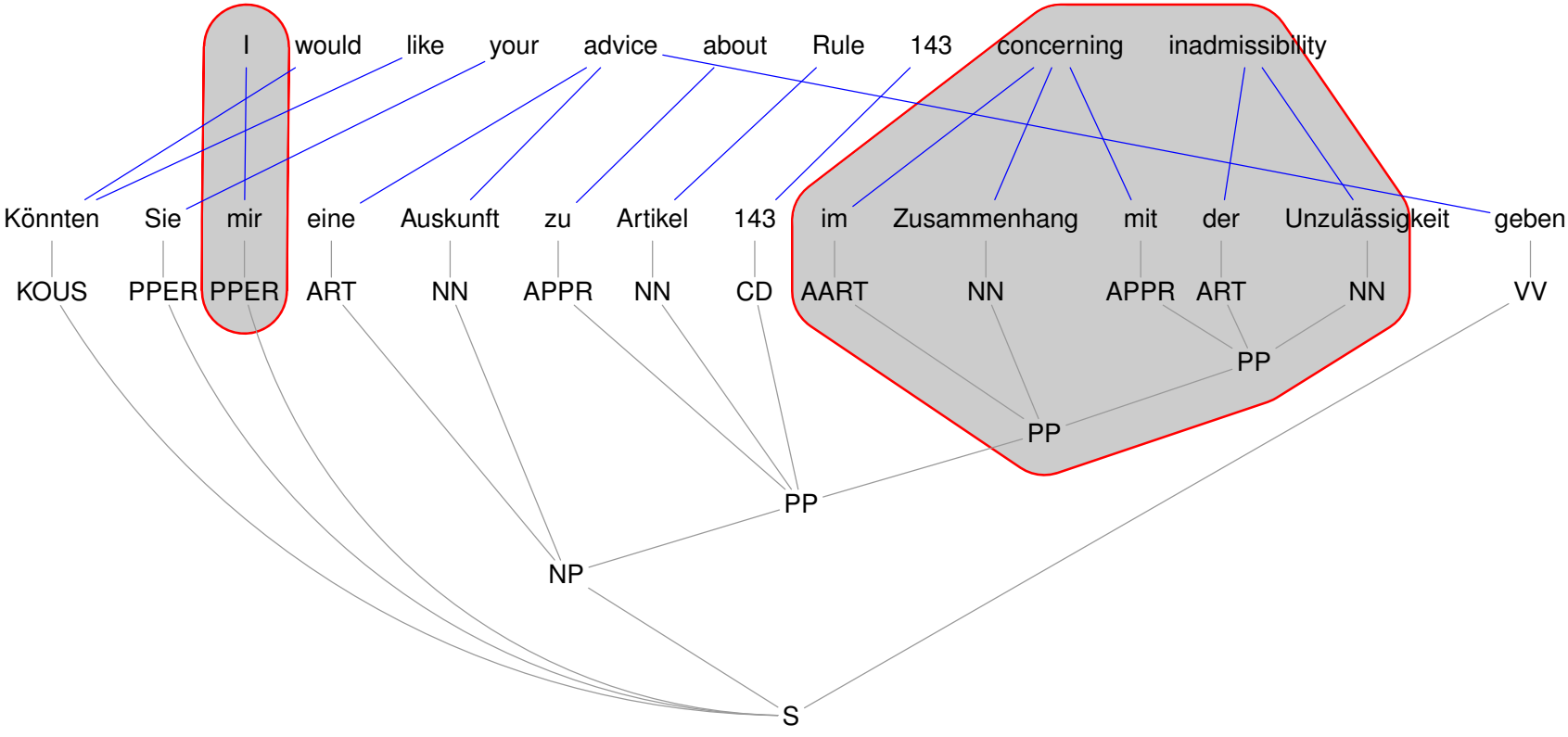
nach [GALLEY, HOPKINS, KNIGHT, MARCU, 2004]



extrahierbare Regeln rot umrandet

Regelextraktion

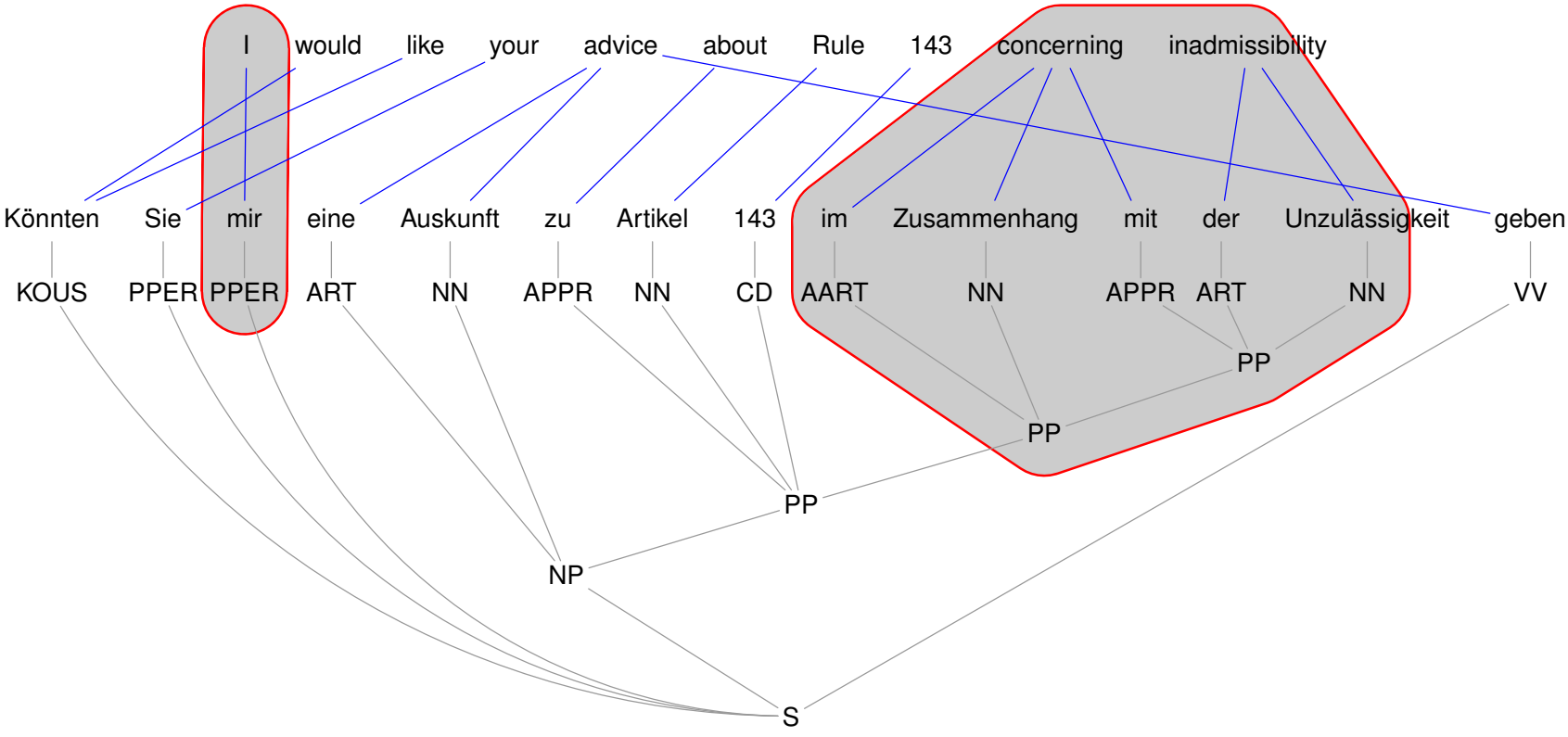
nach [GALLEY, HOPKINS, KNIGHT, MARCU, 2004]



extrahierbare Regeln rot umrandet

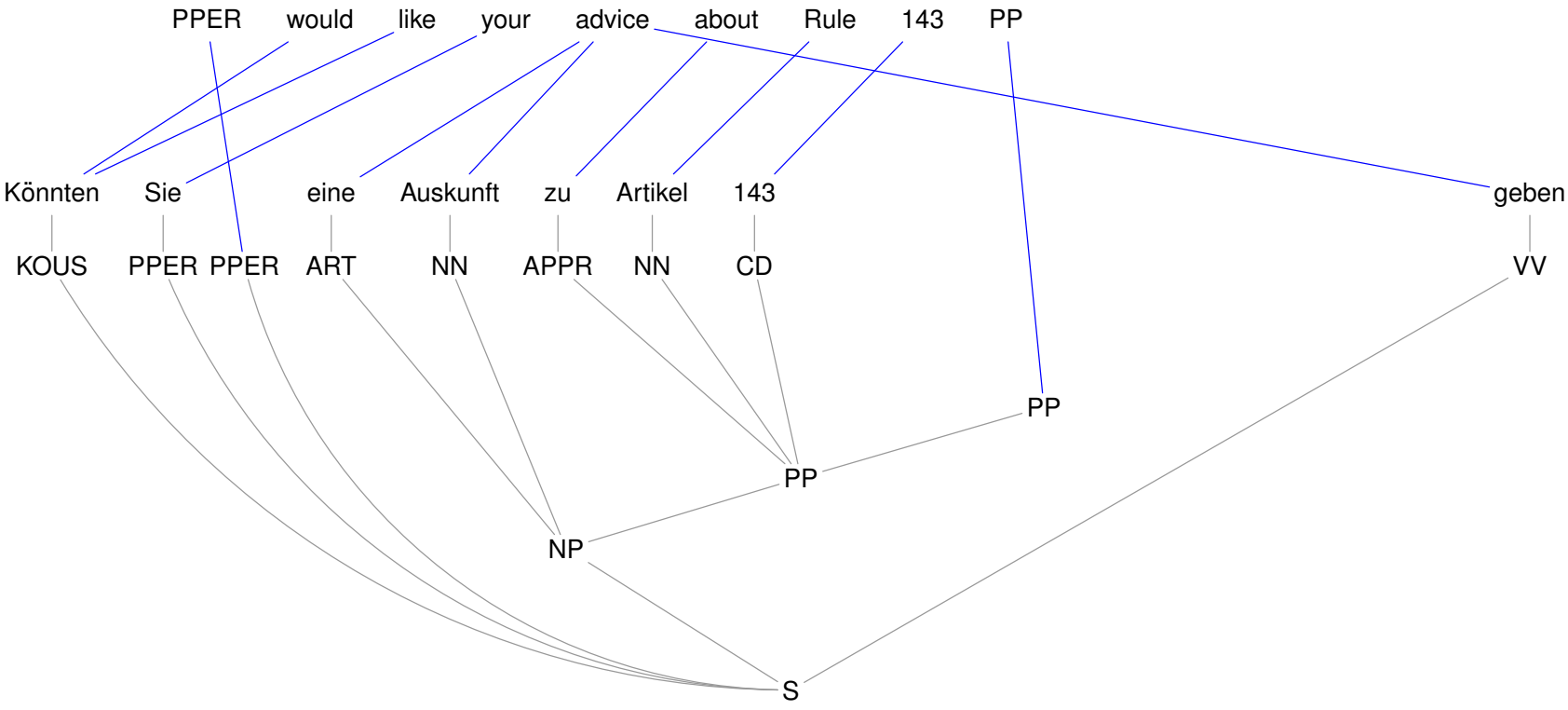
Regelextraktion

Ausschneiden der extrahierbaren Regeln:



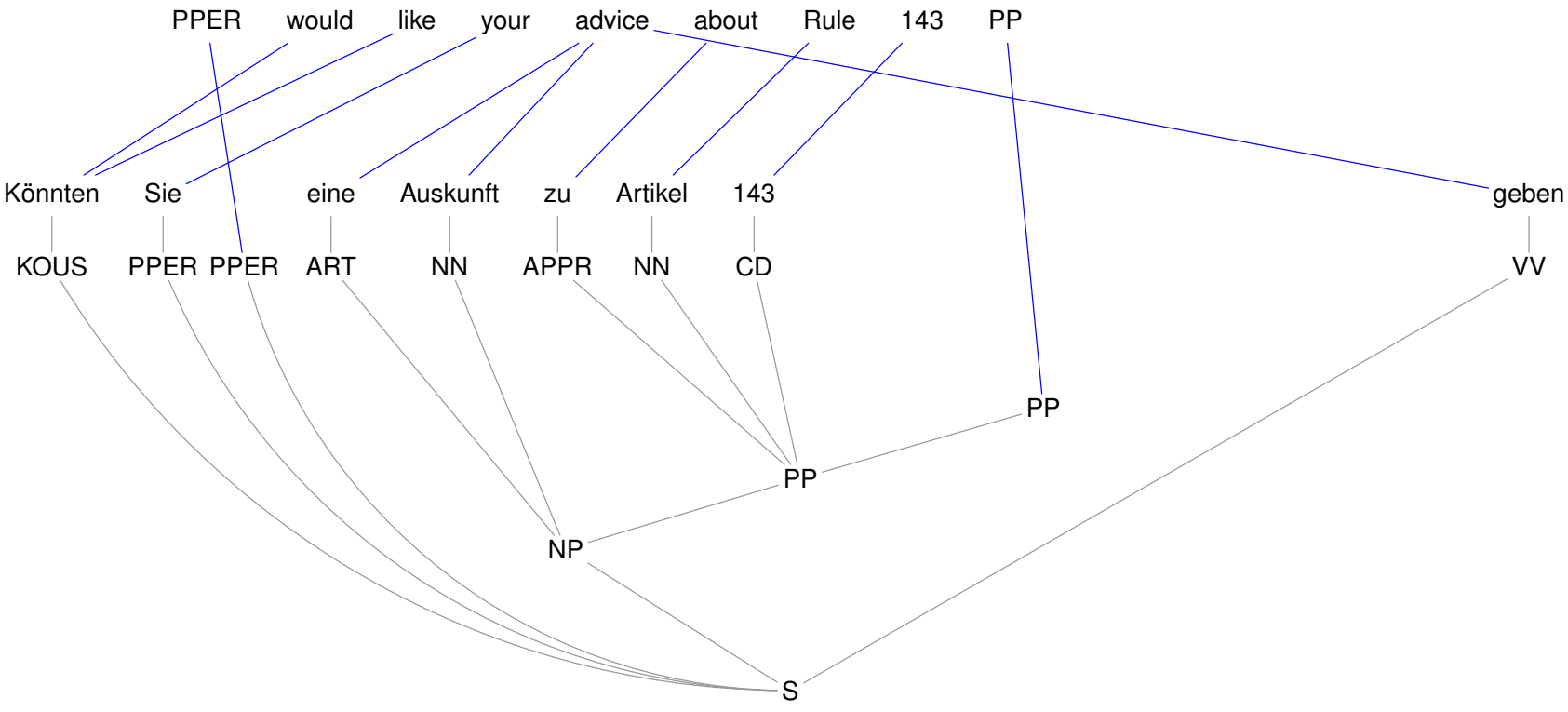
Regelextraktion

Ausschneiden der extrahierbaren Regeln:



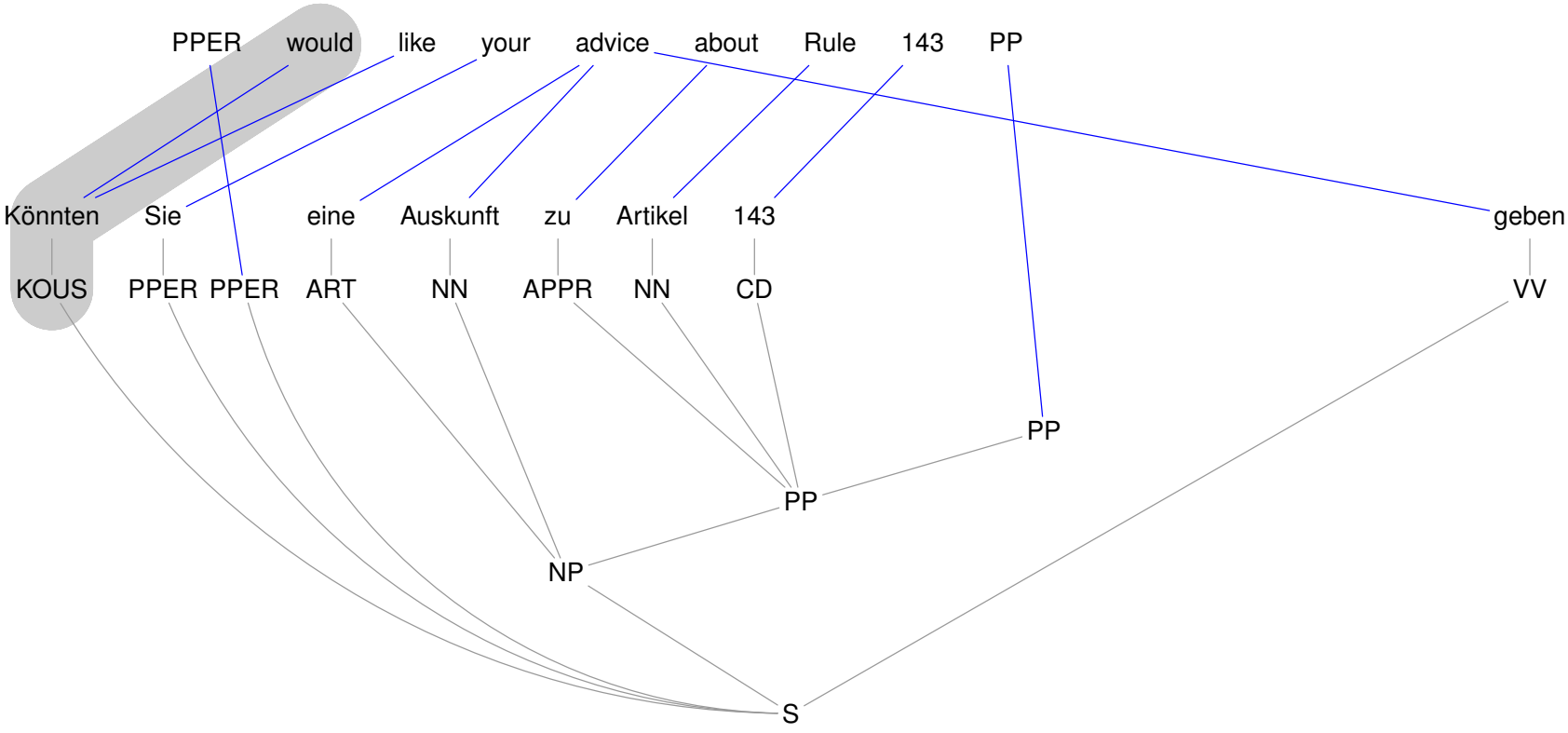
Regelextraktion

Erneute Regelextraktion:



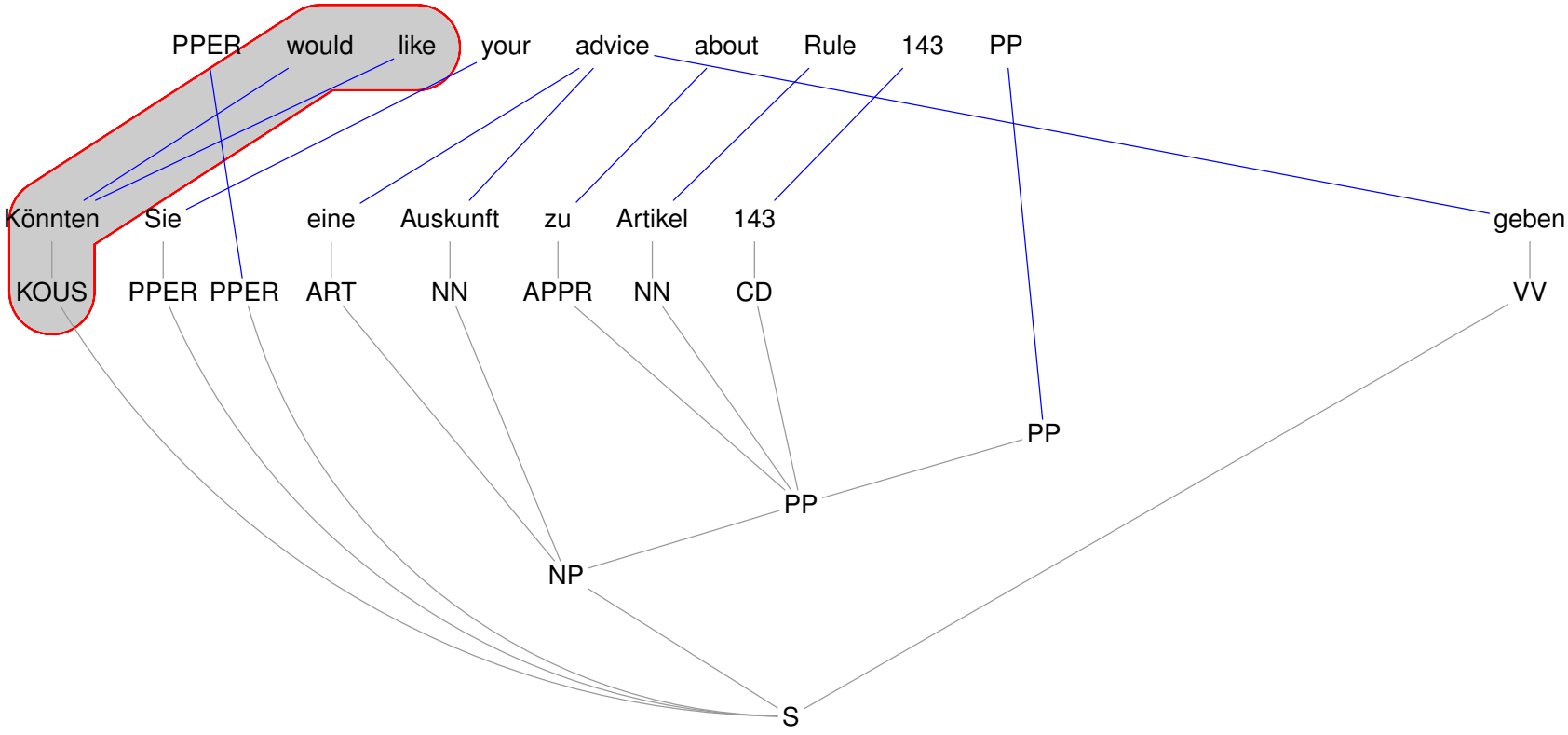
Regelextraktion

Erneute Regelextraktion:



Regelextraktion

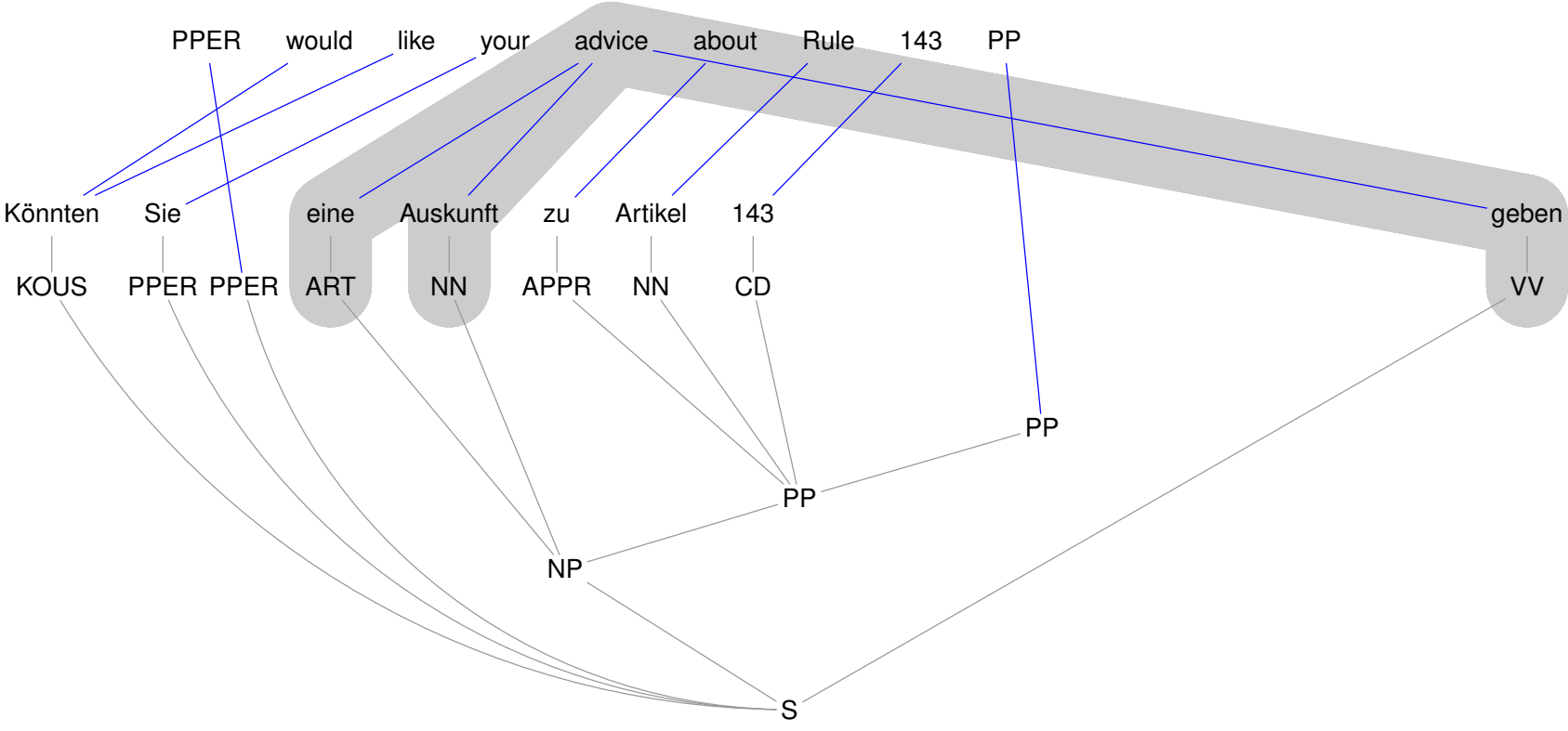
Erneute Regelextraktion:



extrahierbare Regeln rot umrandet

Regelextraktion

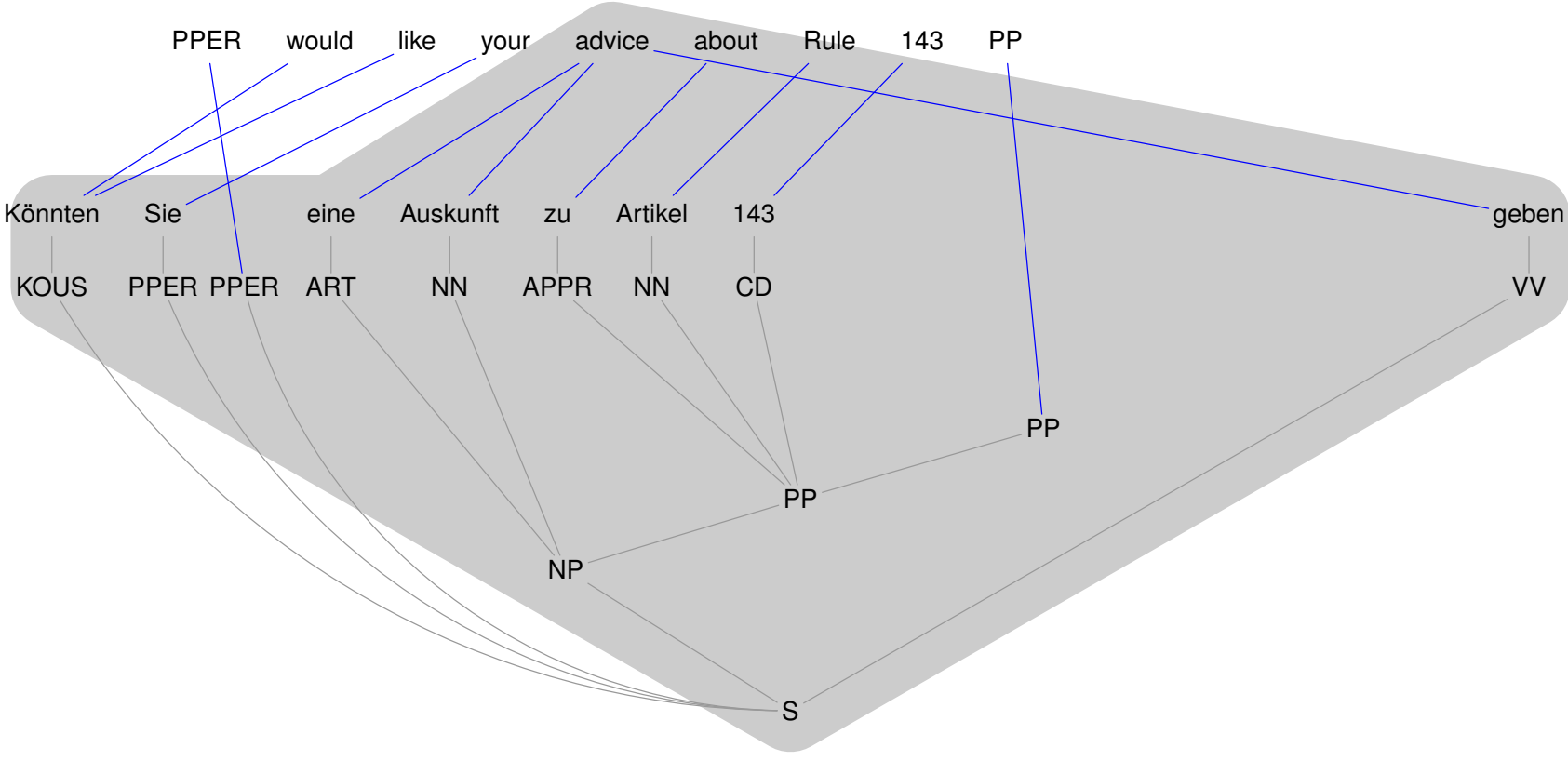
Erneute Regelextraktion:



extrahierbare Regeln rot umrandet

Regelextraktion

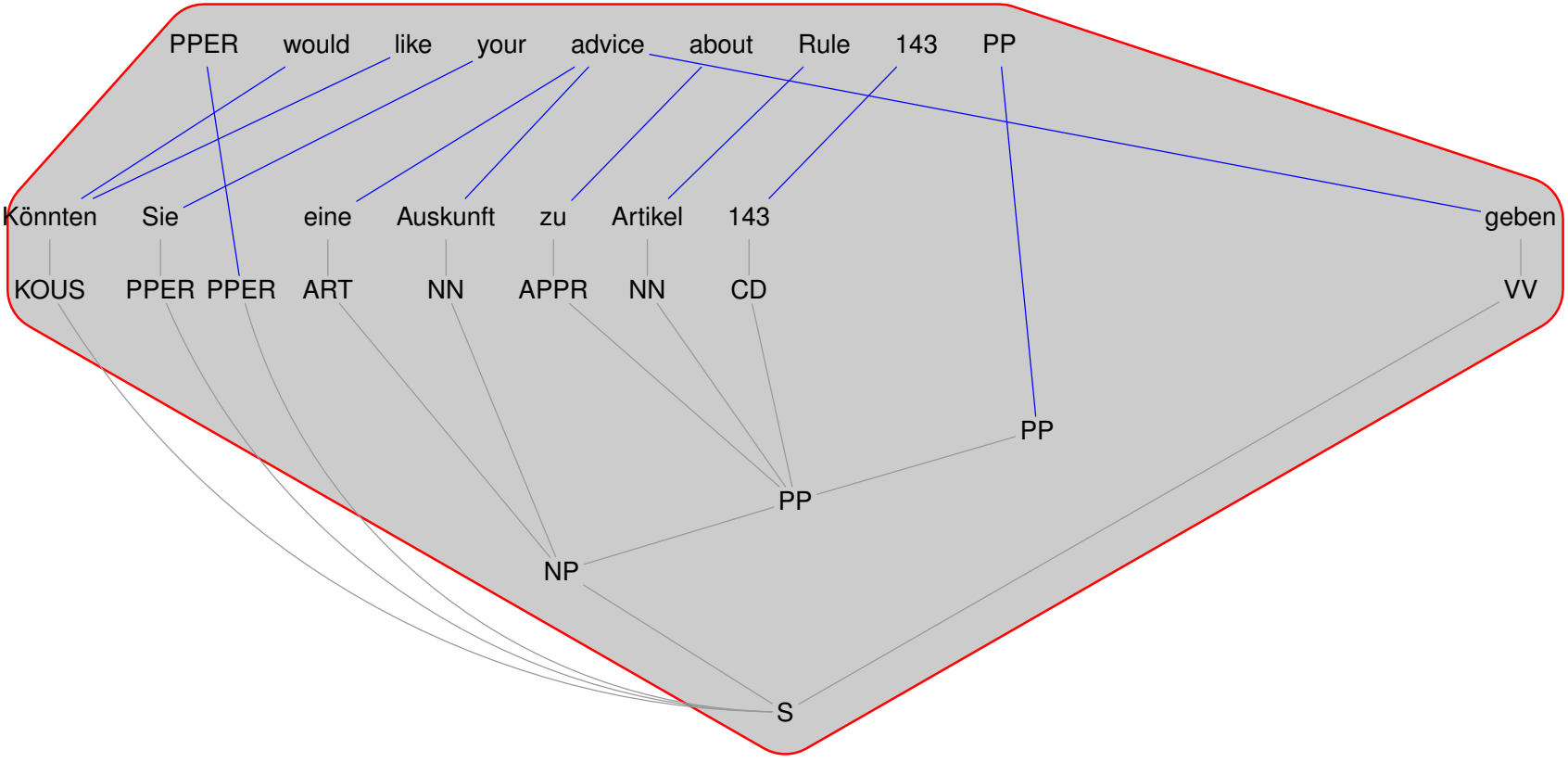
Erneute Regelextraktion:



extrahierbare Regeln rot umrandet

Regelextraktion

Erneute Regelextraktion:



extrahierbare Regeln rot umrandet

Baumübersetzer

Vorteile

- sehr einfach
- implementiert in MOSES [KOEHN et al., 2007]
- “kontext-frei”

Baumübersetzer

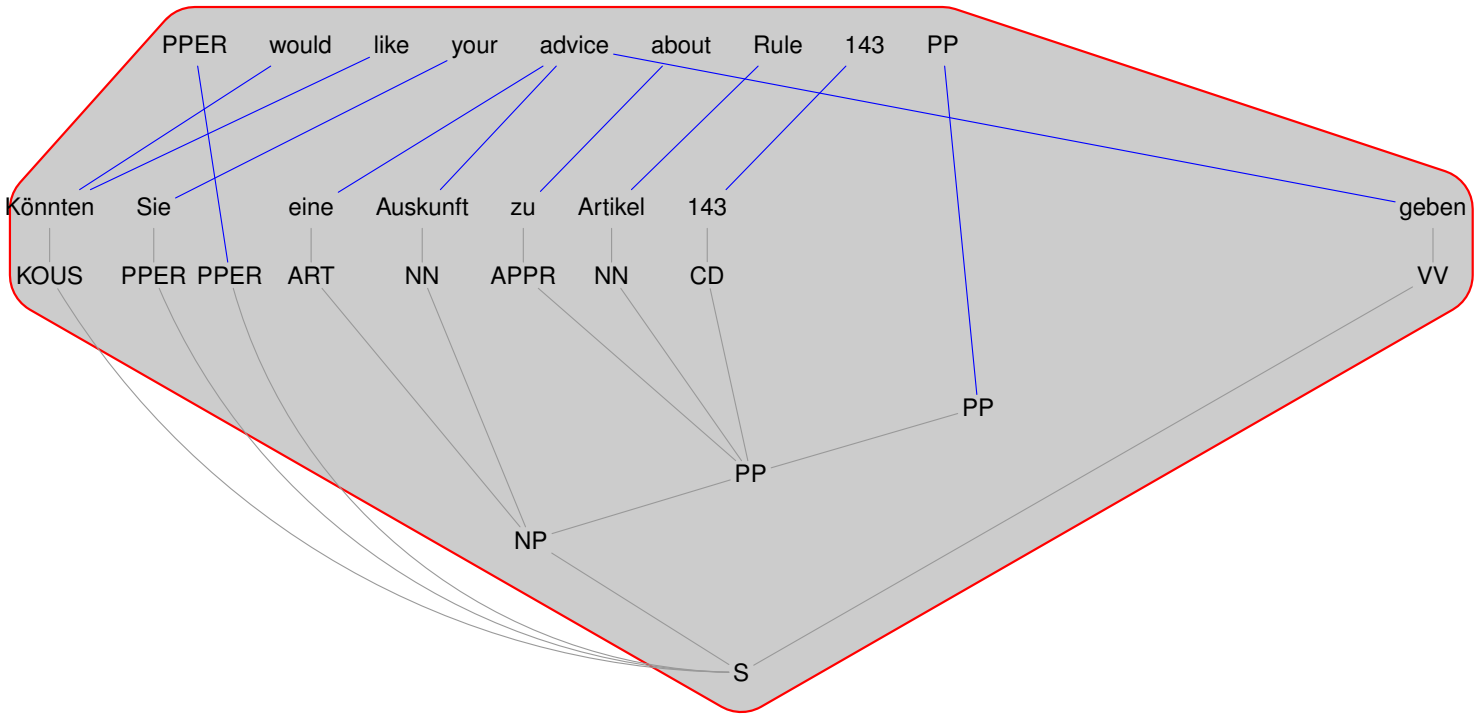
Vorteile

- sehr einfach
- implementiert in MOSES [KOEHN et al., 2007]
- “kontext-frei”

Nachteile

- Probleme mit Diskontinuitäten
- Komposition und Binarisierung problematisch [M. et al., 2009] und [ZHANG et al., 2006]
- “kontext-frei”

Regelextraktion



- sehr spezielle Regel
- jede Regel für “advice” liefert Satzstruktur
- (Syntax hinderlich)

Baumübersetzer

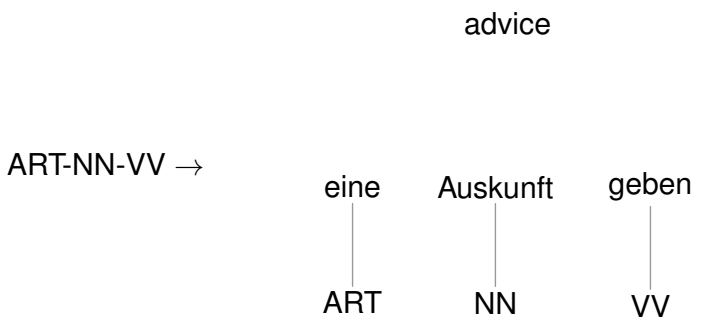
Mehrfach-Baumübersetzer (MBOT)

- Variante von [M., 2010]
- Regeln der Gestalt NT $\rightarrow \left(r, \langle r_1, \ldots, r_n \rangle\right)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseiten $r_1, \ldots, r_n$ einer regulären Baumgr.

Baumübersetzer

Mehrfach-Baumübersetzer (MBOT)

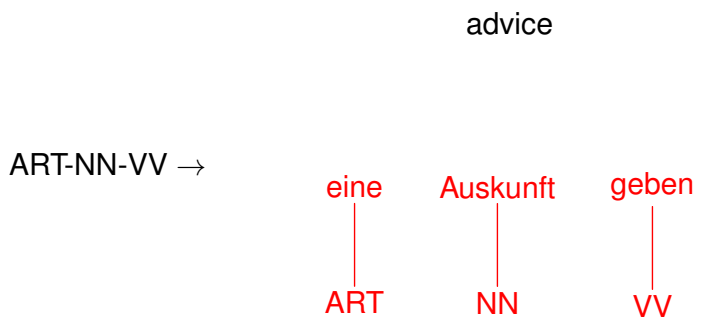
- Variante von [M., 2010]
- Regeln der Gestalt NT $\rightarrow \left(r, \langle r_1, \ldots, r_n \rangle\right)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseiten $r_1, \ldots, r_n$ einer regulären Baumgr.



Baumübersetzer

Mehrfach-Baumübersetzer (MBOT)

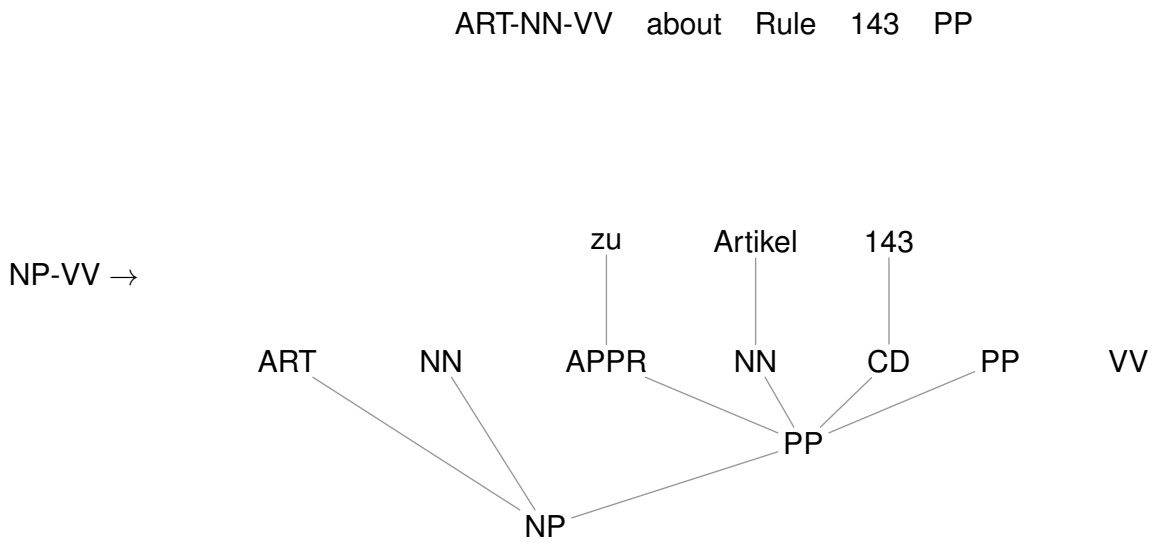
- Variante von [M., 2010]
- Regeln der Gestalt NT $\rightarrow \left(r, \langle r_1, \ldots, r_n \rangle\right)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseiten $r_1, \ldots, r_n$ einer regulären Baumgr.



Baumübersetzer

Mehrfach-Baumübersetzer (MBOT)

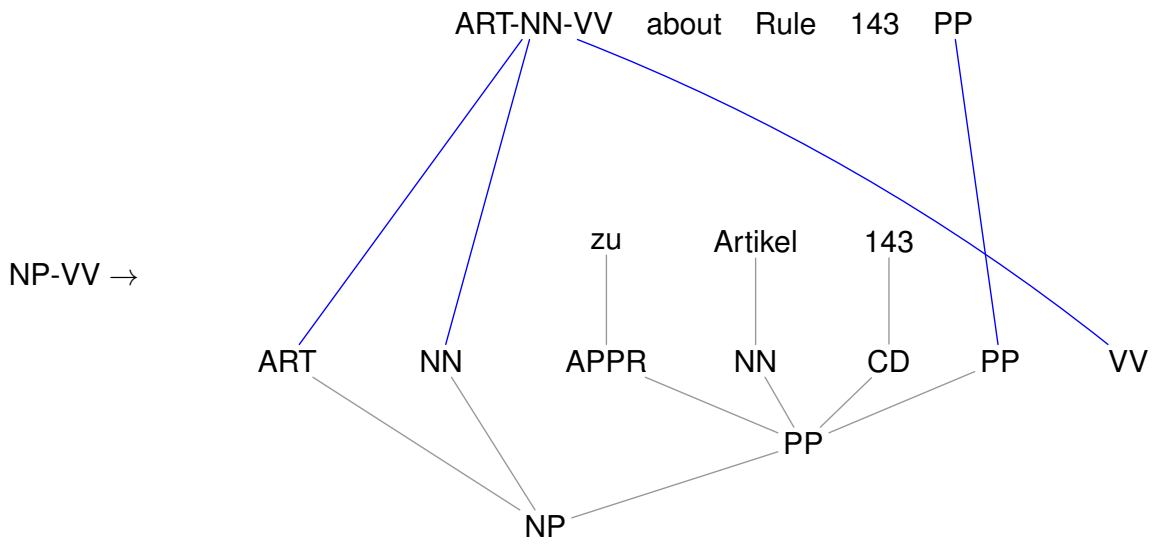
- Variante von [M., 2010]
- Regeln der Gestalt NT $\rightarrow \left(r, \langle r_1, \ldots, r_n \rangle\right)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseiten $r_1, \ldots, r_n$ einer regulären Baumgr.



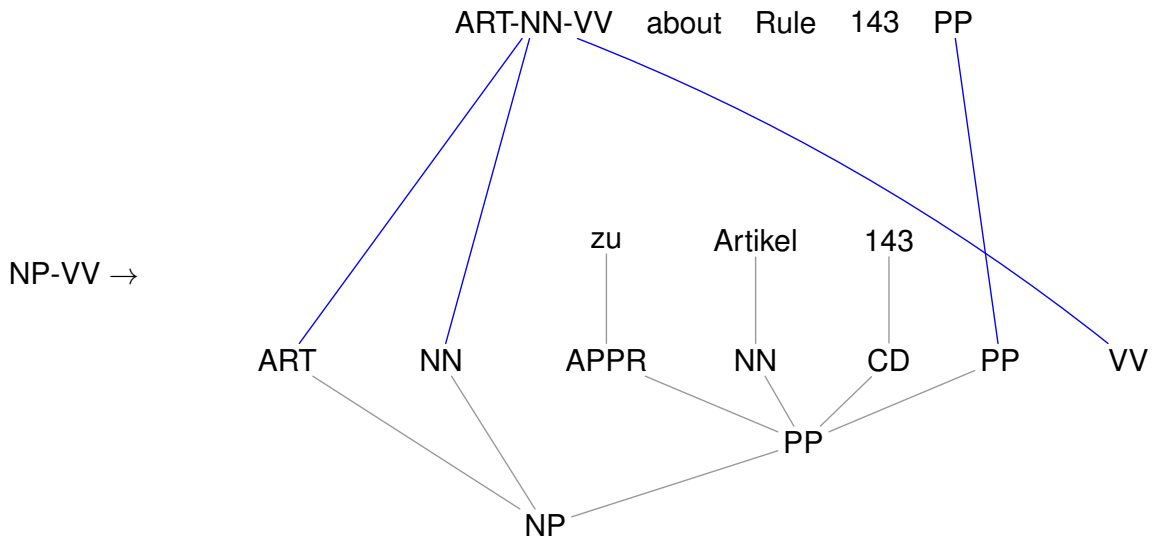
Baumübersetzer

Mehrfach-Baumübersetzer (MBOT)

- Variante von [M., 2010]
- Regeln der Gestalt NT $\rightarrow \left(r, \langle r_1, \ldots, r_n \rangle\right)$
  - Nichtterminal NT
  - rechte Regelseite r einer kontextfreien Grammatik
  - rechte Regelseiten $r_1, \ldots, r_n$ einer regulären Baumgr.
- (surjektive) Synchronisation der Nichtterminale



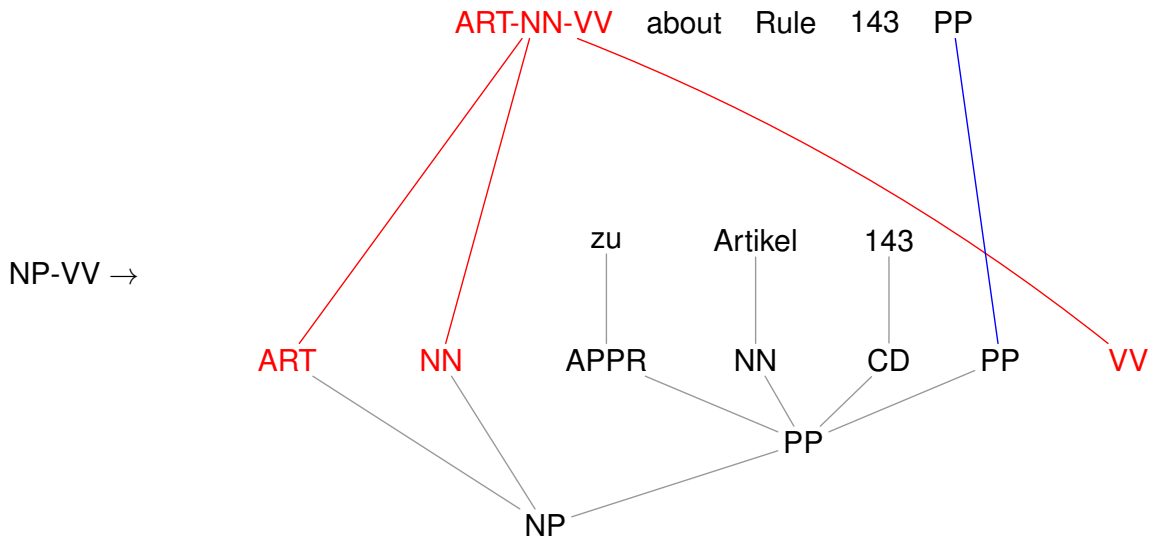
Baumübersetzer



Regelanwendung

1. Nichtterminalauswahl

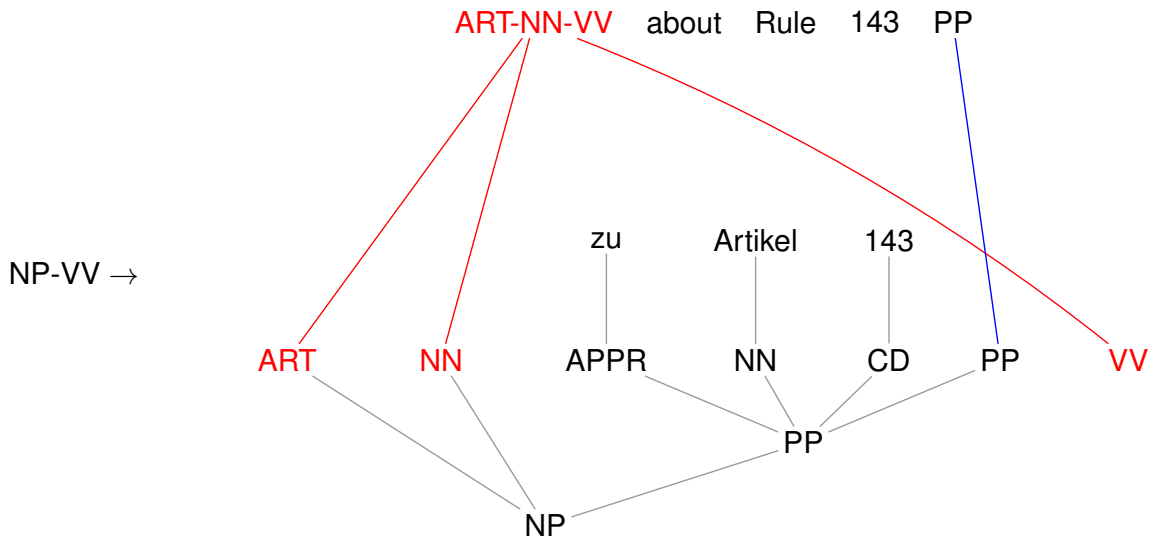
Baumübersetzer



Regelanwendung

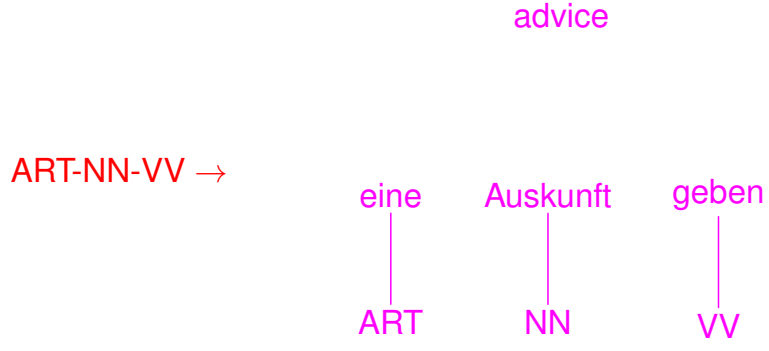
1. Nichtterminalauswahl

Baumübersetzer

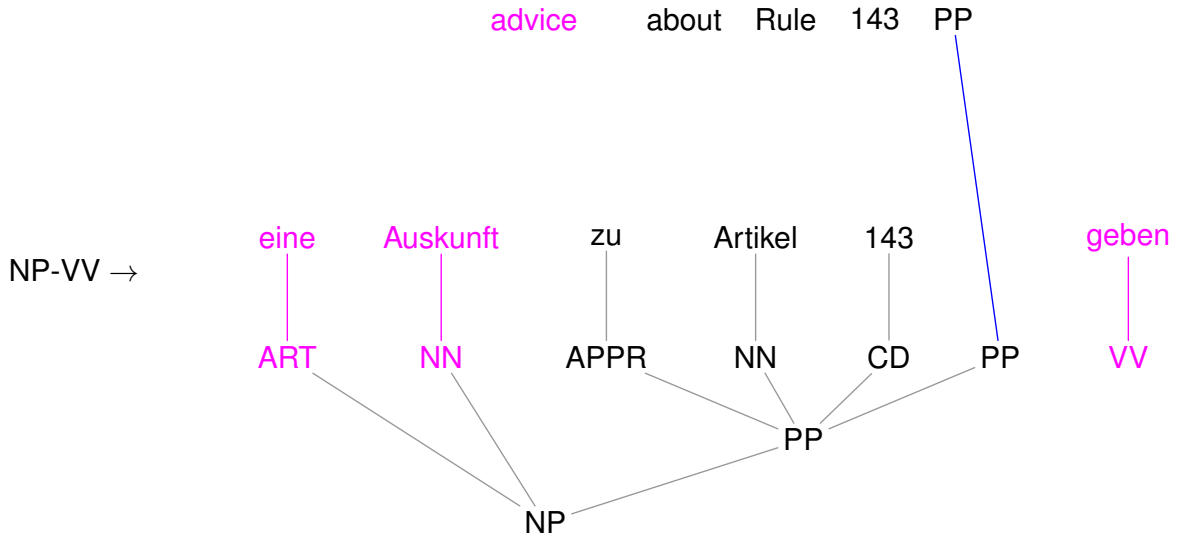


Regelanwendung

1. Nichtterminalauswahl
2. passende Regel

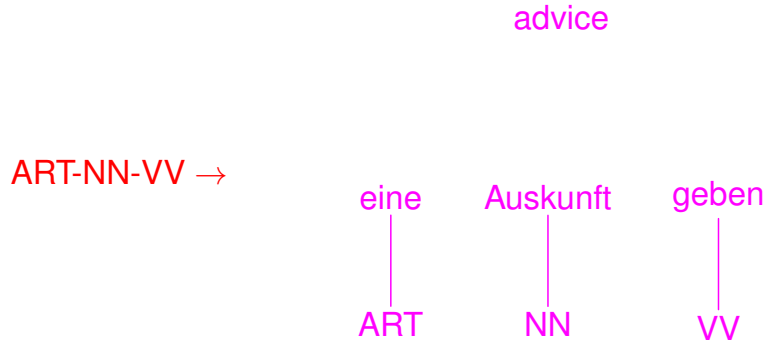


Baumübersetzer



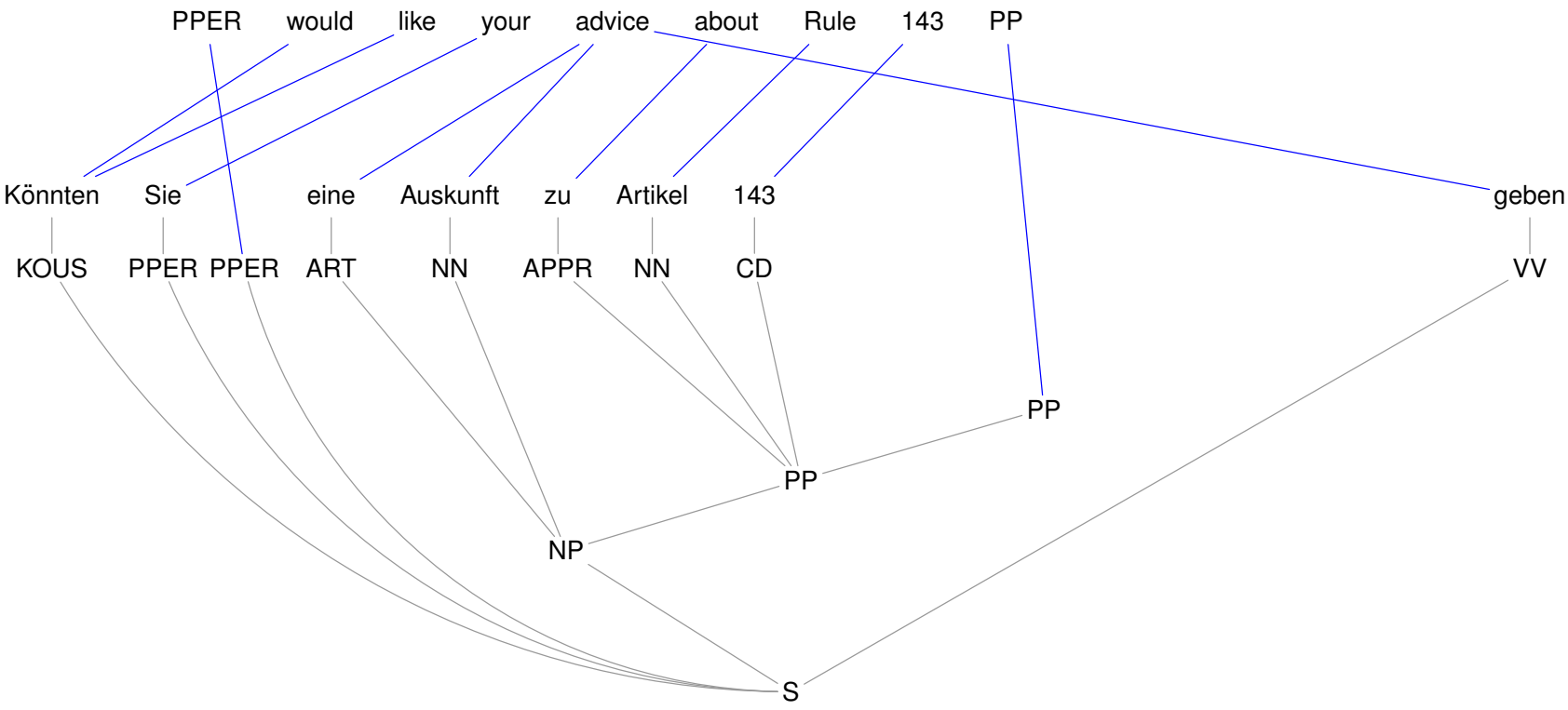
Regelanwendung

1. Nichtterminalauswahl
2. passende Regel
3. Ersetzung



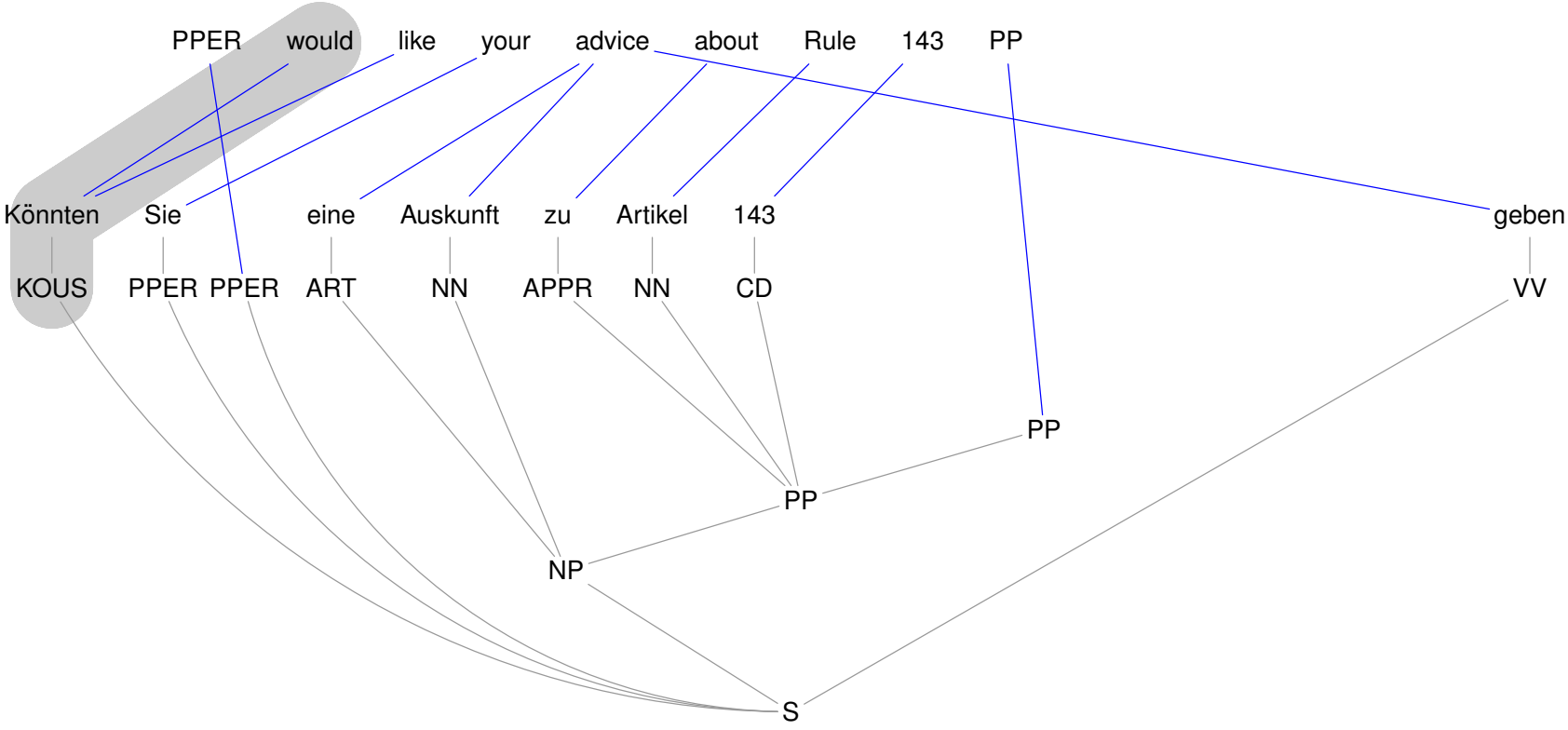
Regelextraktion

nach [M., 2011]



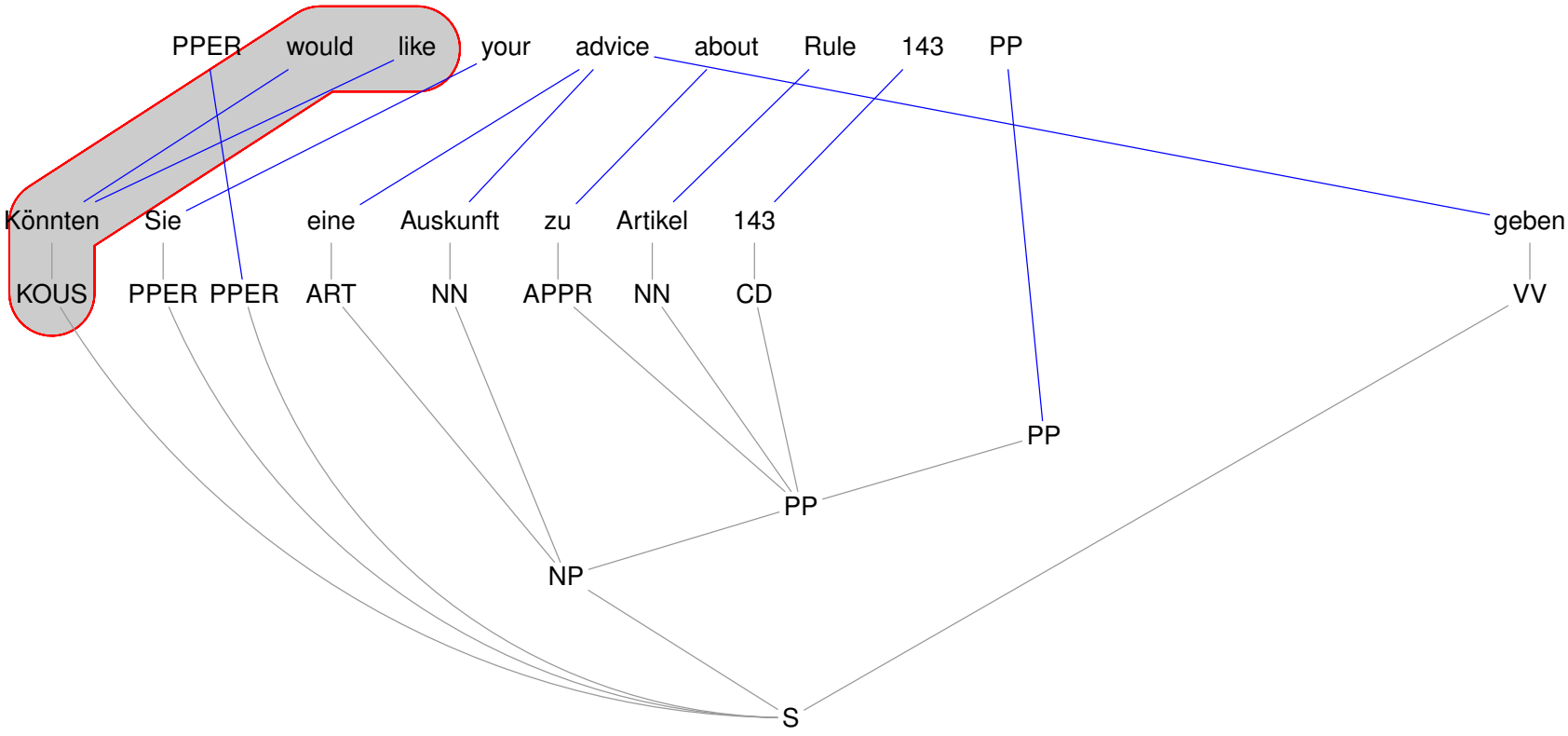
Regelextraktion

nach [M., 2011]



Regelextraktion

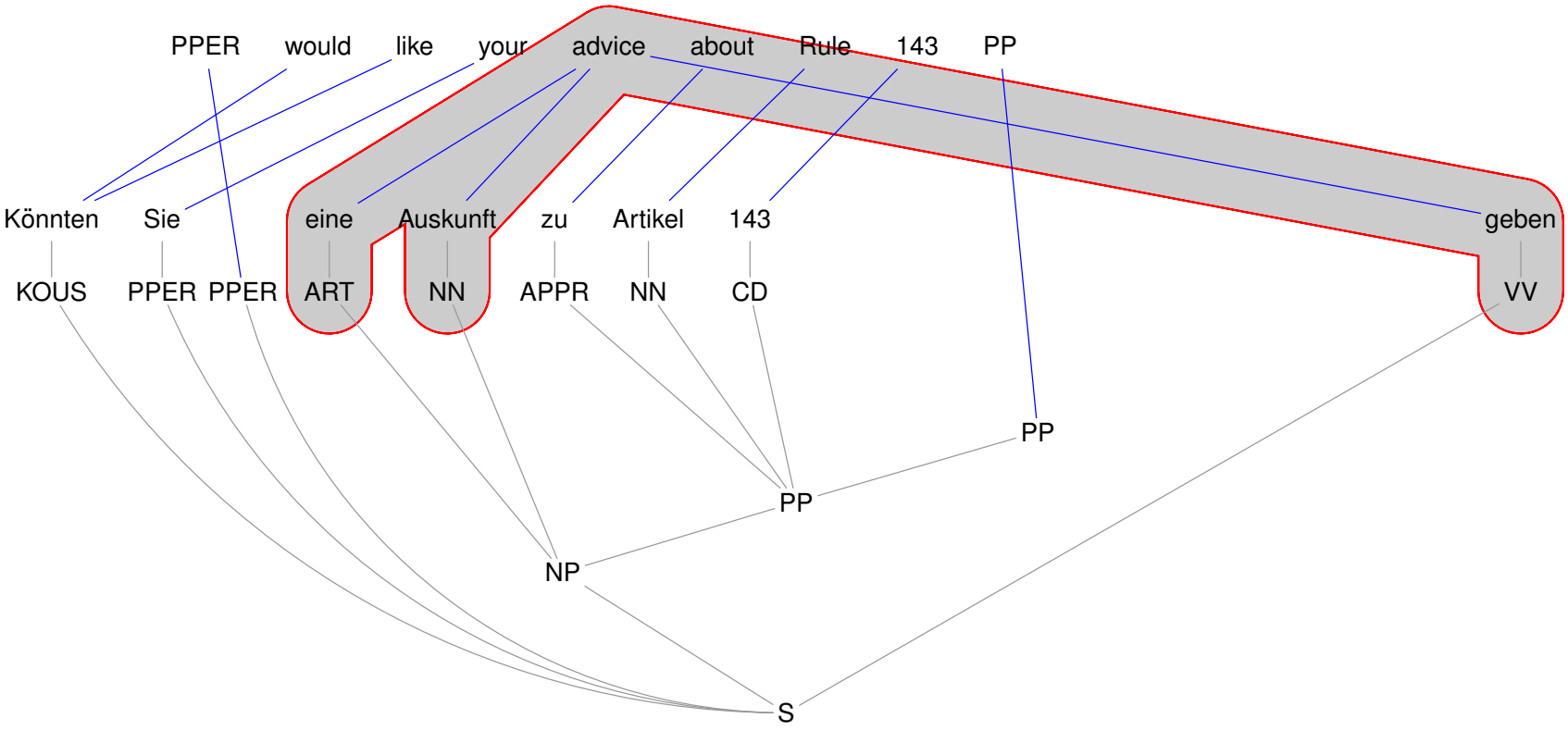
nach [M., 2011]



extrahierbare Regeln rot umrandet

Regelextraktion

nach [M., 2011]



extrahierbare Regeln rot umrandet

Baumübersetzer

Vorteile

- komplizierte Diskontinuitäten
- auch in MOSES [BRAUNE et al., 2013]
- binarisierbar, komponierbar

Baumübersetzer

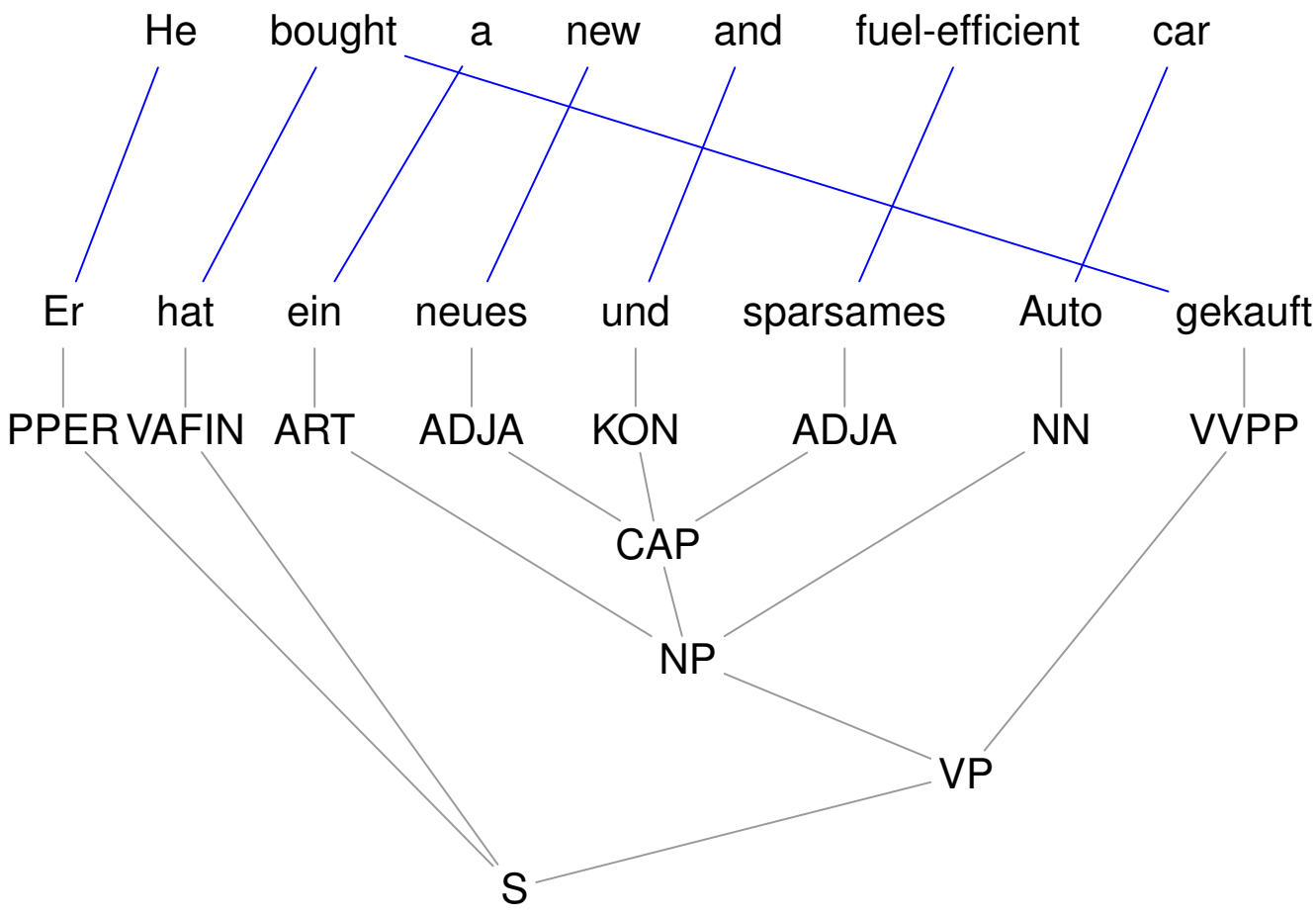
Vorteile

- komplizierte Diskontinuitäten
- auch in MOSES [BRAUNE et al., 2013]
- binarisierbar, komponierbar

Nachteile

- Ausgabe nicht regulär (als Baumsprache)
- nicht symmetrisch (Eingabe kontext-frei; Ausgabe nicht)

Diskontinuitäten



Bewertung

Übersetzung	System	BLEU
Englisch → Deutsch	STSG	15.22
	MBOT	15.90
	phrasenbasiert	16.73
	hierarchisch	16.95
	GHKM	17.10
Englisch → Arabisch	STSG	48.32
	MBOT	49.10
	phrasenbasiert	50.27
	hierarchisch	51.71
	GHKM	46.66
Englisch → Chinesisch	STSG	17.69
	MBOT	18.35
	phrasenbasiert	18.09
	hierarchisch	18.49
	GHKM	18.12

aus [SEEMANN, BRAUNE, M., 2015]

Aktuelle Forschung

Dekodierung

- Eingabeseite kontextfreie Grammatik
- erweiterter CYK-Algorithmus für Übersetzung
  (Parsen des Eingabesatzes; Übersetzung entsteht)

Aktuelle Forschung

Dekodierung

- Eingabeseite kontextfreie Grammatik
- erweiterter CYK-Algorithmus für Übersetzung
  (Parsen des Eingabesatzes; Übersetzung entsteht)

Beobachtungen

- phrasenbasierte System machen keine Suchfehler
  [CHANG, COLLINS, 2011]

Aktuelle Forschung

Dekodierung

- Eingabeseite kontextfreie Grammatik
- erweiterter CYK-Algorithmus für Übersetzung
  (Parsen des Eingabesatzes; Übersetzung entsteht)

Beobachtungen

- phrasenbasierte System machen keine Suchfehler
  [CHANG, COLLINS, 2011]
- STSG und MBOT sehr wohl
  - Heuristiken (??? BLEU)
  - exakte Dekodierung mit Syntaxwald (+2–3 BLEU)

Aktuelle Forschung

Regelextraktion

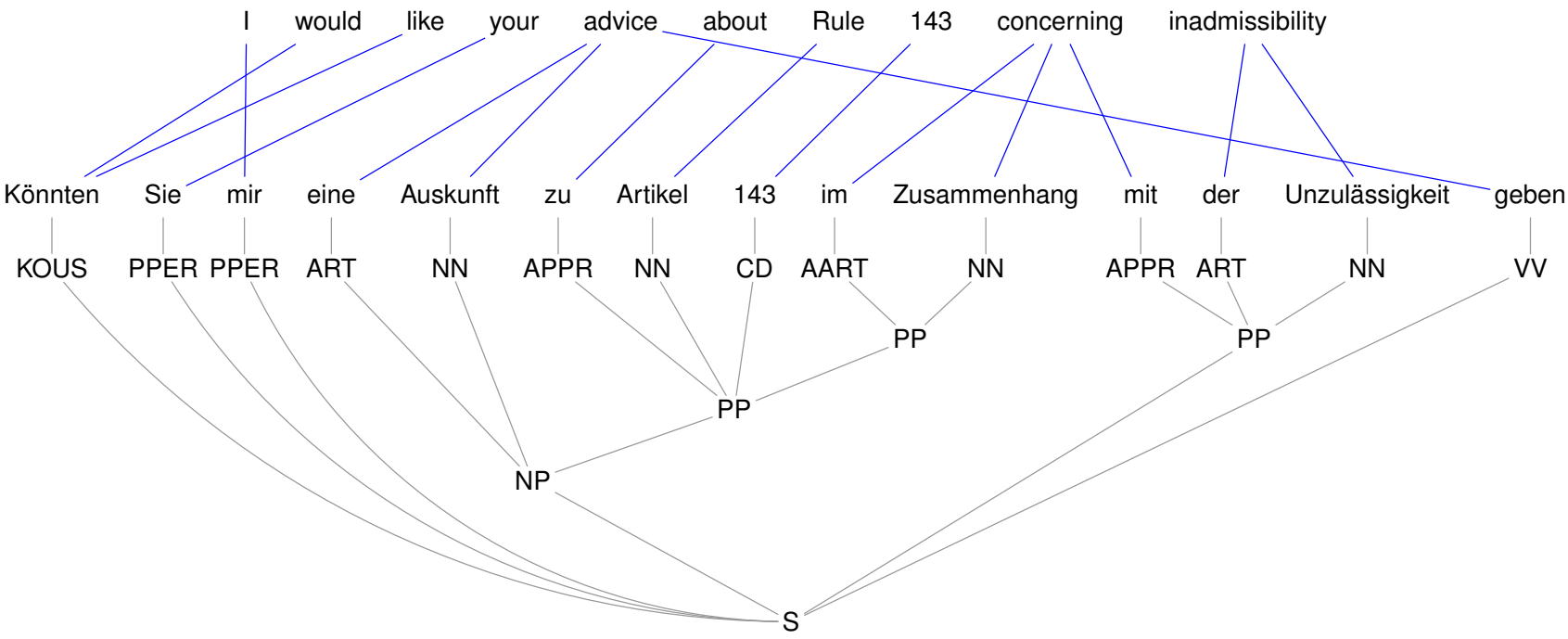
- zu viele extrahierbare Regeln
  - welche Einschränkungen? [SEEMANN, BRAUNE, M., 2015]
  - effizientere Darstellung (evtl. symbolisch)

Aktuelle Forschung

Regelextraktion

- zu viele extrahierbare Regeln
  - welche Einschränkungen? [SEEMANN, BRAUNE, M., 2015]
  - effizientere Darstellung (evtl. symbolisch)
- nur bester Syntaxbaum
  - Regelextraktionen mit Syntaxwald (ambitioniert)

Aktuelle Forschung



Aktuelle Forschung

Regelextraktion

- zu viele extrahierbare Regeln
  - welche Einschränkungen? [SEEMANN, BRAUNE, M., 2015]
  - effizientere Darstellung (evtl. symbolisch)
- nur bester Syntaxbaum
  - Regelextraktionen mit Syntaxwald (ambitioniert)

Übersetzungsmodelle

- nur wortbasierte Systeme für Wortbeziehungen
  - effiziente Einschränkungen moderner Systeme
  - unüberwachtes Lernen

Aktuelle Forschung

Regelextraktion

- zu viele extrahierbare Regeln
  - welche Einschränkungen? [SEEMANN, BRAUNE, M., 2015]
  - effizientere Darstellung (evtl. symbolisch)
- nur bester Syntaxbaum
  - Regelextraktionen mit Syntaxwald (ambitioniert)

Übersetzungsmodelle

- nur wortbasierte Systeme für Wortbeziehungen
  - effiziente Einschränkungen moderner Systeme
  - unüberwachtes Lernen
- Modelle für semantik-basierte Übersetzung
  - graph-basierte Modelle

Aktuelle Forschung

Übersetzungsmodelle

- Ausdrucksstärke
  - von STSG-Kompositionen (mit FÜLÖP, ENGELFRIET)
  - von Teilklassen von MBOT
  - von kombinatorischen Kategorialgrammatiken (mit KUHLMANN)

Aktuelle Forschung

Übersetzungsmodelle

- Ausdrucksstärke
  - von STSG-Kompositionen (mit FÜLÖP, ENGELFRIET)
  - von Teilklassen von MBOT
  - von kombinatorischen Kategorialgrammatiken (mit KUHLMANN)

Minimierung

- klassisch
- basierend auf Simulationen und Bisimulationen
- basierend auf Beinahe-Äquivalenz

Literatur

Auswahl

CHIANG: *Hierarchical Phrase-Based Translation*
Comput. Linguist. 33, 2007

GALLEY, HOPKINS, KNIGHT, MARCU
What's in a Translation Rule? Proc. NAACL, 2004

KOEHN: *Statistical Machine Translation*
Cambridge University Press, 2010

OCH, NEY: *The Alignment Template Approach to Statistical Machine Translation.* Comput. Linguist. 30, 2004