CCG – Statistisches Parsing

Kilian Evang, Christian Wurm

Düsseldorf, 10.11.2022

CCG und statistisches Parsing I

- Hockenmaier and Steedman (2007): CCGbank – semiautomatische Konvertierung der Bäume in der Penn Treebank in CCG-Derivationen
- Clark and Curran (2007): C&C parser – statistischer CCG-Parser mit hoher Genauigkeit
- Honnibal et al. (2010): CCGrebank – CCGbank mit verbesserten NP-Analysen
- Basile et al. (2012): Groningen Meaning Bank – CCG-Baumbank mit semantischer Interpretation, semiautomatisch erstellt mit Hilfe von C&C
- Lewis and Steedman (2014): easyccg – supertagging-basiertes CCG-Parsing (simpler und genau so gut)

CCG und statistisches Parsing II

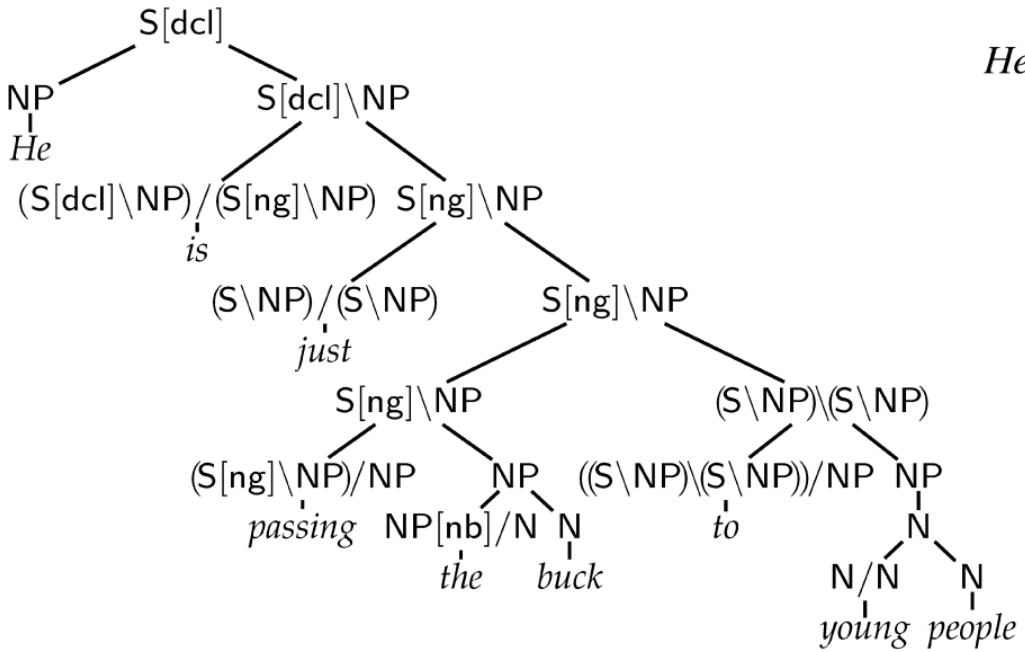
- Evang and Bos (2016); Evang (2016, 2019): CCG-Trainingsdaten für verschiedene Sprachen mit Hilfe von parallelen Korpora automatisch erzeugen
- Abzianidze et al. (2017): Parallel Meaning Bank – CCG-Baumbank mit semantischer Interpretation in vier Sprachen (Englisch, Deutsch, Italienisch, Niederländisch)
- Evang et al. (2019): CCGweb – simples webbasiertes Annotationstool

A PTB Tree

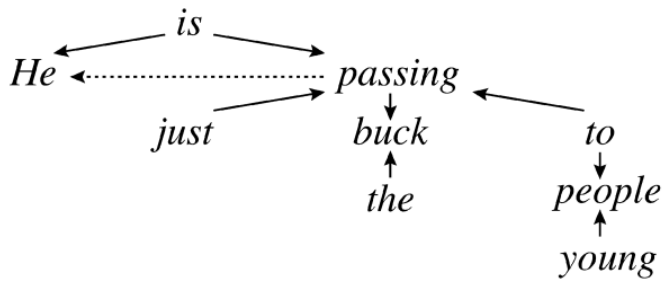
```
(S (NP-SBJ (PRP He))
           (VP (VBZ is)
           (VP (ADVP (RB just))
                (VBG passing)
                (NP (DT the) (NN buck))
                (PP-DIR (TO to)
                     (NP (JJ young) (NNS people)))))
     (. .))
```

Converted to CCG

CCG derivation tree



Dependency graph



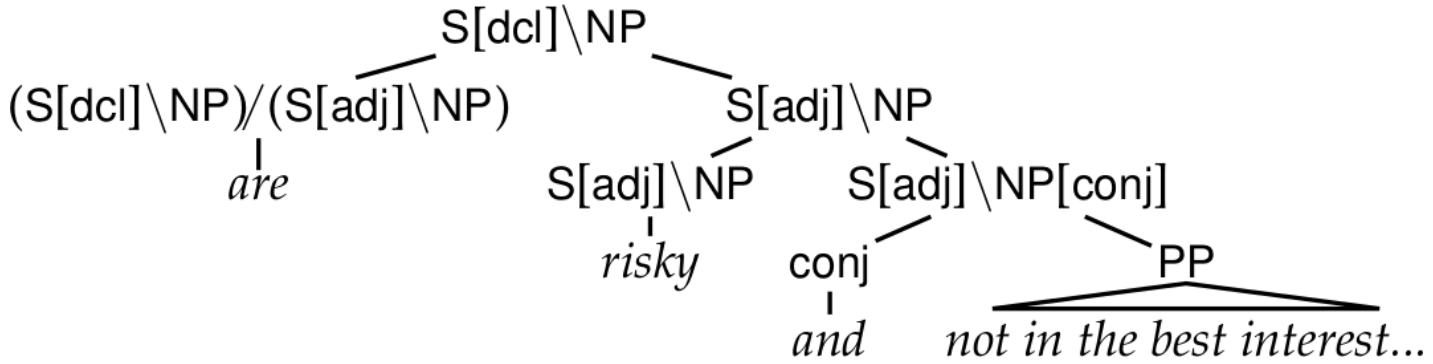
Word-word dependencies

$\langle is,\ (S[dcl]\backslash \mathbf{NP_1})/(S[ng]\backslash NP)_2,\ 1,\ He\rangle$,
$\langle is,\ (S[dcl]\backslash NP_1)/(\mathbf{S[ng]}\backslash \mathbf{NP})_\mathbf{2},\ 2,\ passing\rangle$,
$\langle just,\ (S\backslash NP_1)/(\mathbf{S}\backslash \mathbf{NP})_\mathbf{2},\ 2,\ passing\rangle$,
$\langle passing,\ (S[ng]\backslash \mathbf{NP_1})/NP_2,\ 1,\ He\rangle$,
$\langle passing,\ (S[ng]\backslash NP_1)/\mathbf{NP_2},\ 2,\ buck\rangle$,
$\langle the,\ NP[nb]/\mathbf{N_1},\ 1,\ buck\rangle$,
$\langle to,\ ((S\backslash NP_1)/(\mathbf{S}\backslash \mathbf{NP})_\mathbf{2})/NP_3,\ 2,\ passing\rangle$,
$\langle to,\ ((S\backslash NP_1)/(S\backslash NP)_2)/\mathbf{NP_3},\ 3,\ people\rangle$,
$\langle young,\ N/\mathbf{N_1},\ 1, people\rangle$

Another PTB Tree

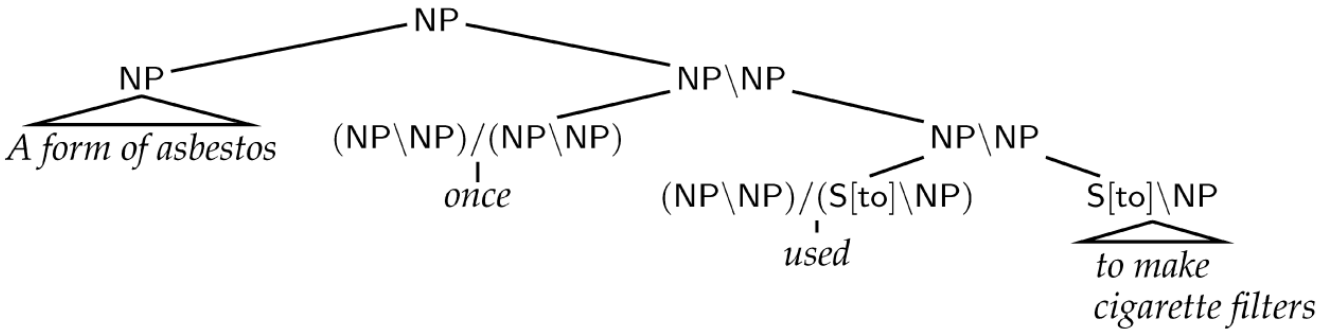
```
(VP (VBP are)
    (UCP-PRD (ADJP risky)
             (CC and)
             (PP not in the best interest of the investing public)))
```

Converted to CCG

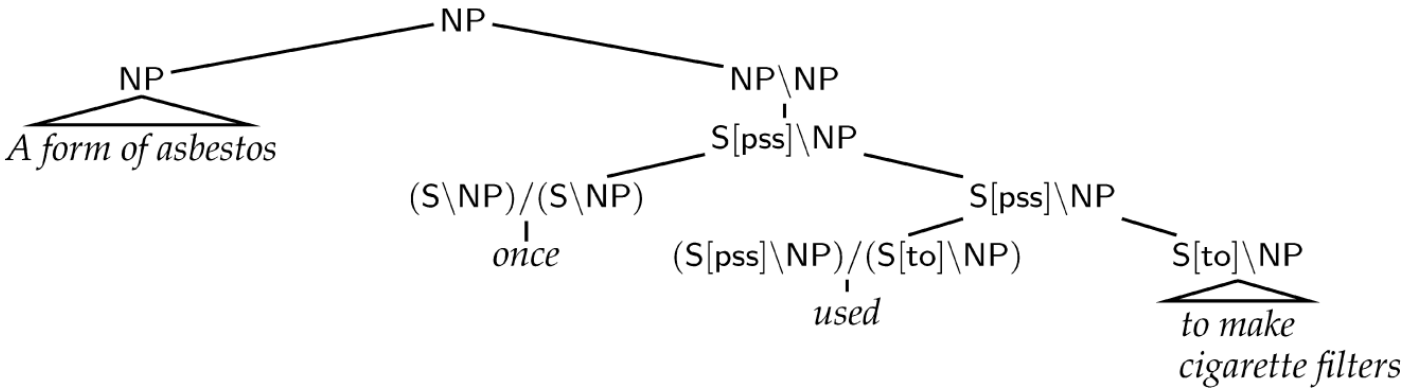


Type Changing I

a) Standard CCG derivation:



b) CCG derivation with type-changing rules:



Type Changing II

a. $S[pss]\backslash NP_i \Rightarrow NP_i\backslash NP_i$
"workers [exposed to it]"

b. $S[adj]\backslash NP_i \Rightarrow NP_i\backslash NP_i$
"a forum [likely to bring attention to the problem]"

c. $S[ng]\backslash NP_i \Rightarrow NP_i\backslash NP_i$
"signboards [advertising imported cigarettes]"

d. $S[ng]\backslash NP_i \Rightarrow (S\backslash NP_i)\backslash(S\backslash NP_i)$
"become chairman, [succeeding Ian Butler]"

e. $S[dcl]/NP_i \Rightarrow NP_i\backslash NP_i$
"the millions of dollars [it generates]"

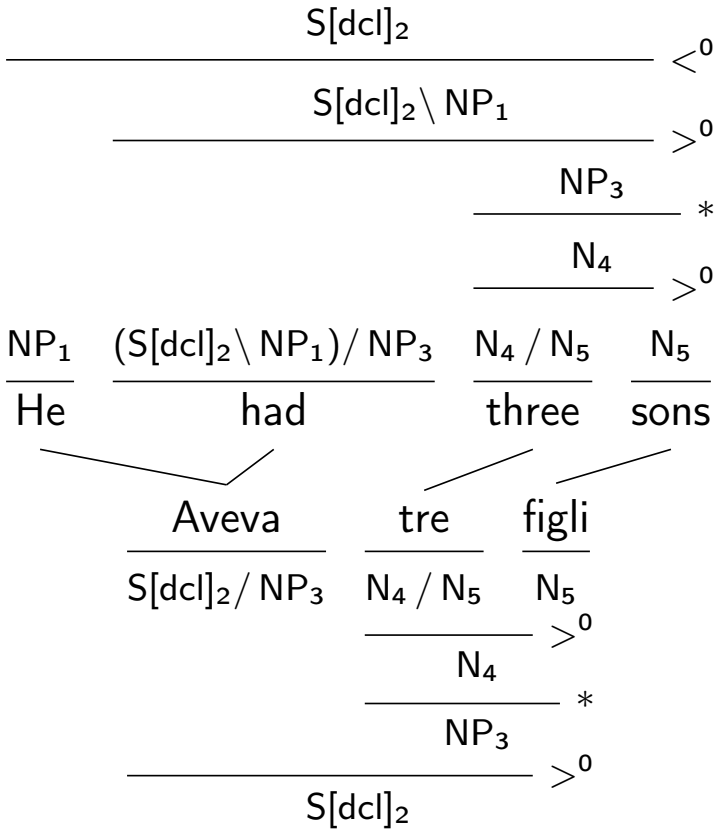
CCGbank-Annotationsschema

- S-Typen: S[dcl], S[q] für Ja-Nein-Fragen, S[wh] für W-Fragen, S[emb] für Komplementsätze, S[pt] für *past participle*, S[ng] für *ing*-Form...
- Modifizierer (Adjunkte) unterspezifiziert für Typ
- Attributive Adjektive sind N-Modifizierer (N / N), prädikative Adjektive sind ein Art VP (S[adj]\ NP)
- Binär verzweigende Koordination
- Type Changing
- Spezialregeln für Satzzeichen

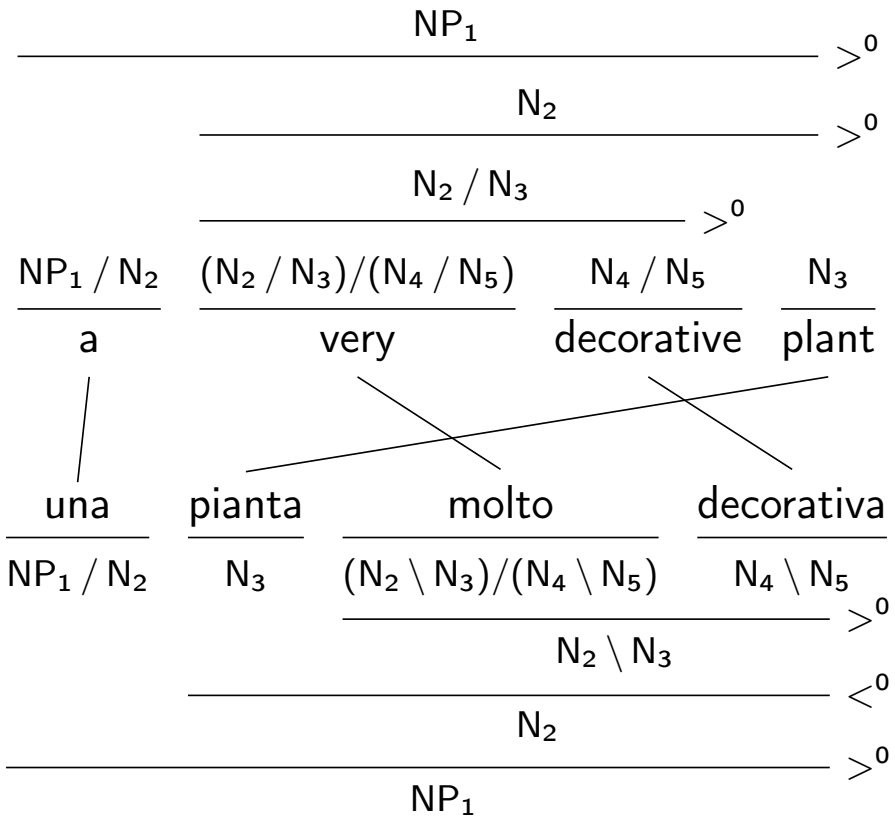
Cross-lingual CCG Induction

- Was, wenn wir eine neue Sprache mit CCG parsen wollen, aber keine Traningsdaten haben?
- Wir generieren uns welche.
- Idee: parallele Daten English/Zielsprache
- Englische Sätze parsen
- Automatisches Word Alignment
- Kategorien vom Englischen in die Zielsprache „projizieren“
- Methode hat Grenzen: Quellparses nicht perfekt, Word Alignments nicht perfekt, funktioniert nur bei strukturerhaltenden Übersetzungen, Features und Type-Changing-Regeln für Englisch nicht unbedingt geeignet für andere Sprachen
- Aber: Ergebnis als Vorannotation brauchbar

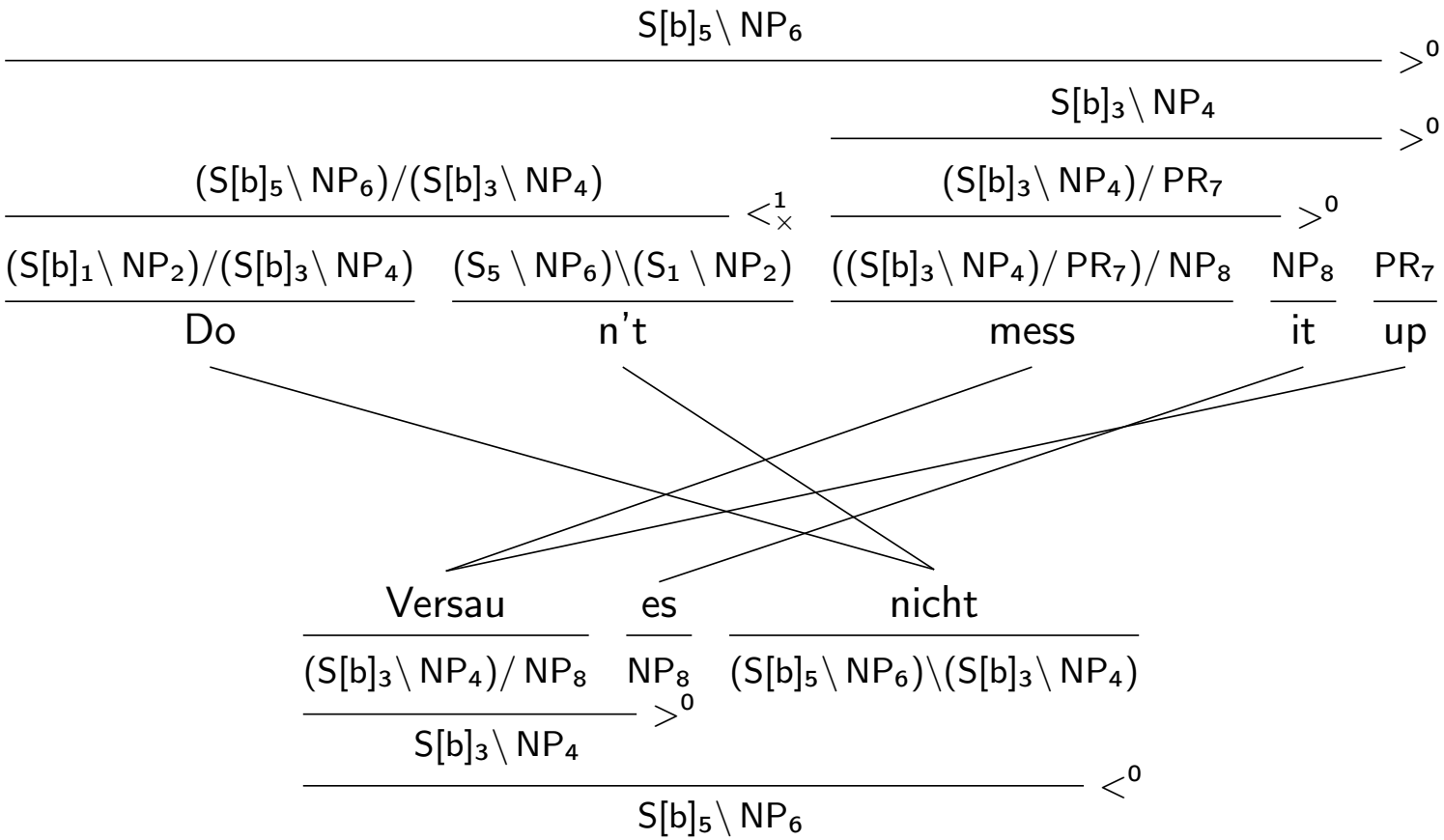
Beispiel 1/3


S[dcl]2
<0
S[dcl]2\NP1
>0
NP3
*
N4
>0
NP1
(S[dcl]2\NP1)/NP3
N4/N5
N5
He
had
three
sons
Aveva
tre
figli
S[dcl]2/NP3
N4/N5
N5
>0
N4
*
NP3
>0
S[dcl]2

Beispiel 2/3


NP1
N2
N2 / N3
NP1 / N2
(N2 / N3)/(N4 / N5)
N4 / N5
N3
a
very
decorative
plant
una
pianta
molto
decorativa
NP1 / N2
N3
(N2 \ N3)/(N4 \ N5)
N4 \ N5
N2 \ N3
N2
NP1
>0
>0
>0
>0
<0
>0

Beispiel 3/3


S[b]5 \ NP6
>0
S[b]3 \ NP4
>0
(S[b]5 \ NP6)/(S[b]3 \ NP4)
<1×
(S[b]3 \ NP4)/ PR7
>0
(S[b]1 \ NP2)/(S[b]3 \ NP4)
(S5 \ NP6)\(S1 \ NP2)
((S[b]3 \ NP4)/ PR7)/ NP8
NP8
PR7
Do
n't
mess
it
up
Versau
es
nicht
(S[b]3 \ NP4)/ NP8
NP8
(S[b]5 \ NP6)\(S[b]3 \ NP4)
>0
S[b]3 \ NP4
<0
S[b]5 \ NP6

Praktische Annotationsübung

https://ccgweb.phil.hhu.de/

CCGweb-Annotationsschema

Ähnlich wie CCGbank, aber

- Satzzeichen und Koordination verwenden normale Applikationsregeln
- Genitivkonstruktionen werden mit PP analysiert
- Handbuch: `https://ccgweb.phil.hhu.de/manual.php`

Zusammenfassung

Kategorialgrammatiken sind superspannend!

Literatur I

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3196–3200, Istanbul, Turkey. European Language Resources Association (ELRA).

Literatur II

Clark, S. and Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Evang, K. (2016). *Cross-lingual learning of an Open-domain Semantic Parser*. PhD thesis, University of Groningen.

Evang, K. (2019). Cross-lingual CCG induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1577–1587, Minneapolis, Minnesota. Association for Computational Linguistics.

Literatur III

Evang, K., Abzianidze, L., and Bos, J. (2019). CCGweb: a new annotation tool and a first quadrilingual CCG treebank. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Evang, K. and Bos, J. (2016). Cross-lingual learning of an open-domain semantic parser. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 579–588, Osaka, Japan. The COLING 2016 Organizing Committee.

Hockenmaier, J. and Steedman, M. (2007). CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Literatur IV

Honnibal, M., Curran, J. R., and Bos, J. (2010). Rebanking CCGbank for improved NP interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 207–215, Uppsala, Sweden. Association for Computational Linguistics.

Lewis, M. and Steedman, M. (2014). A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar. Association for Computational Linguistics.