# Learning linearization rules from treebanks

**Geert-Jan Kruijff**

**NEGRA EM6/SFB 378**

*Computational Linguistics*

*Saarland University*

*Saarbrücken, Germany*

⟨GJ@COLI.UNI-SB.DE⟩

# Goal

- **Use a treebank to learn rules describing linearization**
  - Restrictions on linearization
  - Variability in linearization
- **Work in progress**

- **Acknowledgments**

Oliver Čulo (SAM/HMMr), Christian Korthals, Ivana Kruijff-Korbayová, Richard Moot (UU:CGN), Shravan Vasishth (R), Zdeněk Žabokrtsky (CU:PDT)

This work is supported by the DFG Sonderforschungsbereich 378 *Resource-Sensitive Cognitive Processes*, Project NEGRA EM6.

# Why learn rules describing linearization?

- **Practical issues**
  - *Treebank grammar*: CFG productions read off TB trees (Charniak, 1996)
  - ⇒ Is a CFG strong enough?
  - ⇒ Coverage, generalizability with freer word order?
- **Theoretical issues**
  - Cross-linguistic primitives
  - Hierarchy of generative strength or performance complexity

# Overview

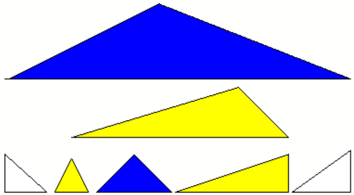
- **Four Type X languages** (Greenberg, 1966; Hawkins, 1983)
  - English, German, Dutch, Czech
  - Different degrees of word order freedom (Steele, 1978; Kruijff, 2001)
- **Empirical investigations**
  - Discontinuous wordgroups (more freedom ⇒ more discontinuity)
  - Scrambling (more freedom ⇒ more scrambling)

  ⇒ Beyond CFG, range of variability
- **Learning linearization rules**
  - Learning problem
  - Approach: General setting, intuitions
  - First experiments
- **Conclusions & outlook**

# Investigation: Discontinuous wordgroups

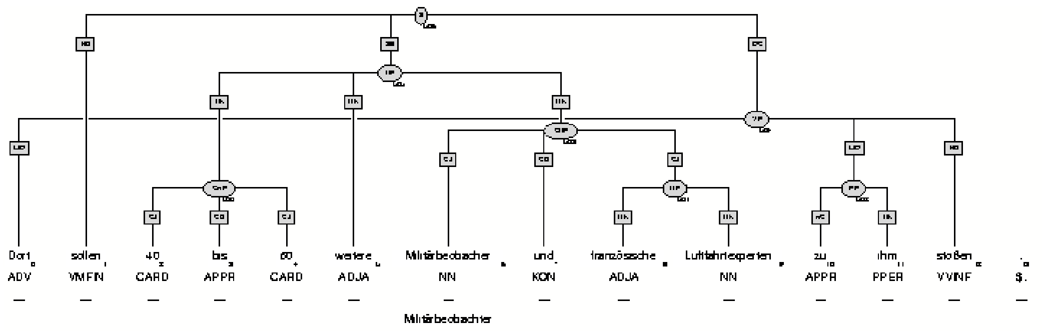
- **Discontinuity in wordgroups**
  - Intervening material from non-dependent wordgroups
  - ⇒ Discontinuous span
- **Null hypothesis**: **More freedom ⇒ more discontinuity**

# Discontinuous wordgroups (once)



- Yellow span interrupted once: intervening blue material

# Discontinuous wordgroups (once)

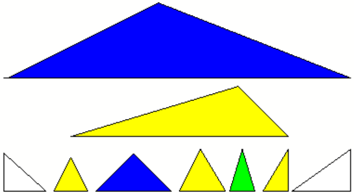



"Dort sollen 40 bis 50 weitere Militärbeobachter und französische Luftfahrtexperten zu ihm stoßen."

(*There, another 40 to 50 military observers and French airline experts should push towards him.*)

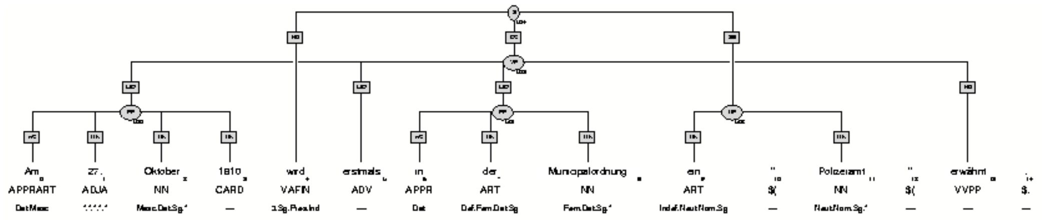
[NEGRA 14876]

# Discontinuous wordgroups (twice)



- Yellow span interrupted twice: intervening blue and green material

# Discontinuous wordgroups (twice)




“Am 27. Oktober 1810 wird erstmals in der Municipalordnung ein Polizeiamt erwähnt.”

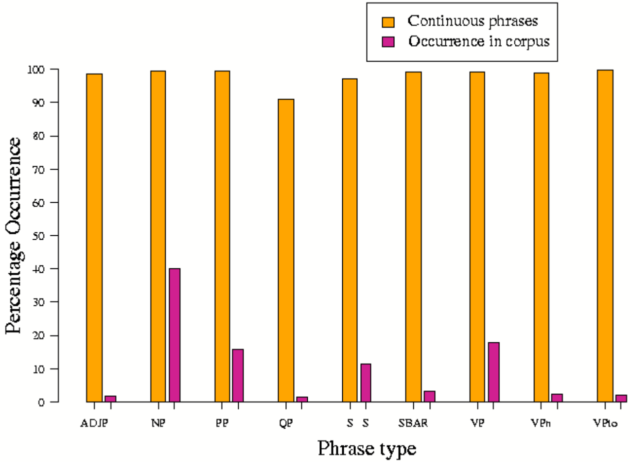
(*On the 27th of October 1810 a police office is mentioned for the first time in the municipality.”*)

[NEGRA 165]

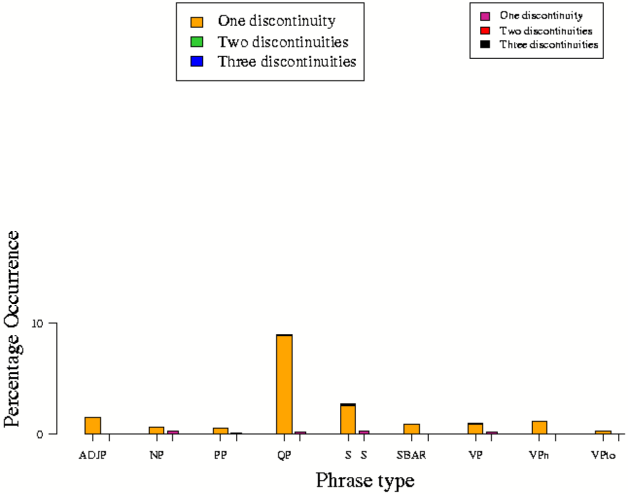
# Experimental setup

- **Four treebanks**
  - English: PTB WSJ section, 49208 trees; (et al, 1995)
  - German: NEGRA, 20602 trees (Skut et al., 1997)
  - Dutch: CGN, 29571 trees (Oostdijk, 2000)
  - Czech: PDT Row 5, 21992 trees (Hajič, 1998)
- **Extract modification context for each head**
  - Dependency tree
  - Includes both complements and adjuncts
- **Count (dis)continuity**
  - Number of interruptions of a head's span
  - 0=continuous, >0 discontinuous
  - Relative freq. (#occ. of head), absolute freq. (#observed nodes)
  - Presentation: Phrases with frequency ≥1%, up to 3 holes (≥0.01%)

# Results: Continuous wordgroups in English (WSJ)



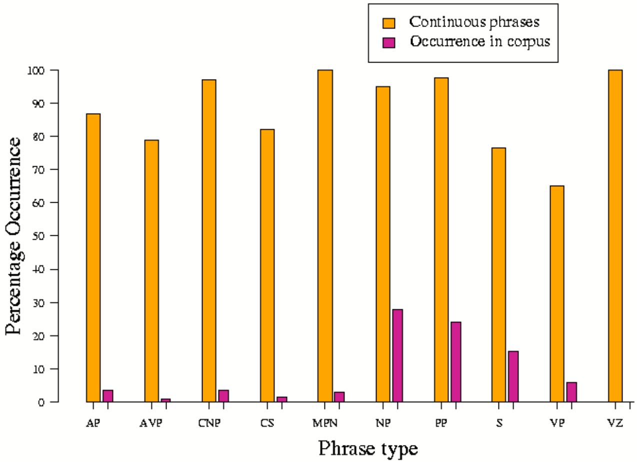
# Results: Discontinuous wordgroups in English (WSJ)



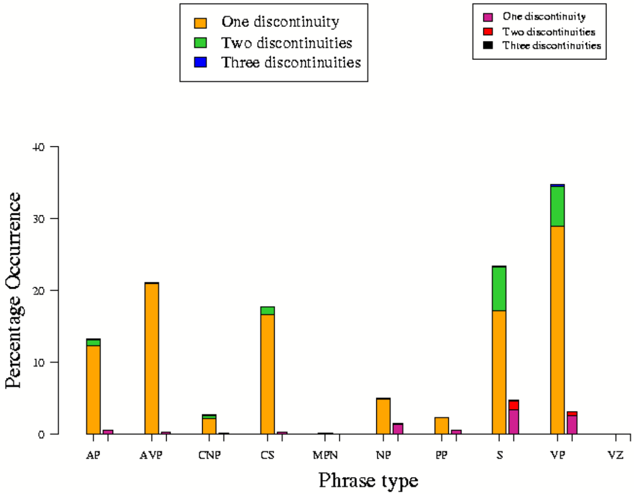
# Frequency (in %) of (dis)continuous occ. per total # occ. (WSJ)

| **Type** | **0 holes** | **1 hole** | **2 holes** | **3 holes** | $\Sigma$ |
|---|---|---|---|---|---|
| ADJP | 1.64% (98.54%) | 0.02% | 0.00% | 0.00% | 0.02% |
| NP | 40.13% (99.39%) | 0.24% | 0.00% | 0.00% | 0.24% |
| PP | 15.69% (99.51%) | 0.08% | 0.00% | 0.00% | 0.08% |
| QP | 1.41% (91.06%) | 0.14% | 0.00% | 0.00% | 0.14% |
| S | 11.56% (97.26%) | 0.31% | 0.00% | 0.00% | 0.31% |
| SBAR | 3.32% (99.11%) | 0.03% | 0.00% | 0.00% | 0.03% |
| VP | 17.99% (99.06%) | 0.16% | 0.01% | 0.00% | 0.17% |
| VPn | 2.47% (98.84%) | 0.03% | 0.00% | 0.00% | 0.03% |
| VPto | 2.17% (99.71%) | 0.01% | 0.00% | 0.00% | 0.01% |
| | $\Sigma^0 = 96.38\%$ | $\Sigma^1 = 1.02\%$ | $\Sigma^2 = 0.01\%$ | $\Sigma^3 = 0.00\%$ | $\Sigma^d = 1.03\%$ |

# Results: Continuous wordgroups in German (NEGRA)



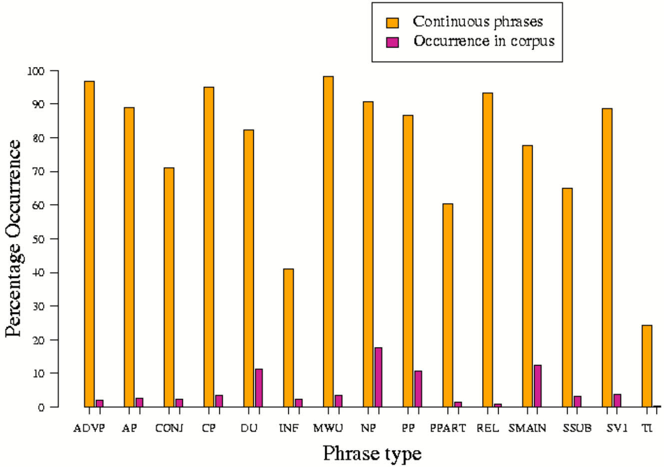
# Results: Discontinuous wordgroups in German (NEGRA)



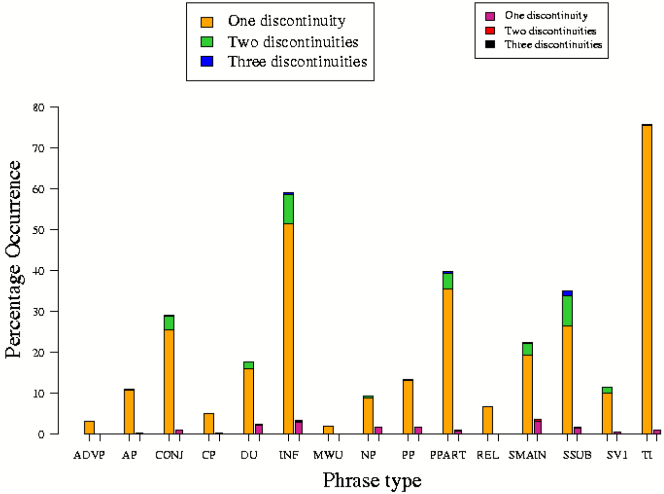
# Frequency (in %) of (dis)continuous occ. per total # occ. (NEGRA)

| **Type** | **0 holes** | **1 hole** | **2 holes** | **3 holes** | $\Sigma$ |
|---|---|---|---|---|---|
| AP | 3.52% (86.80%) | 0.50% | 0.04% | 0.00% | 0.54% |
| AVP | 1.01% (78.90%) | 0.27% | 0.00% | 0.00% | 0.27% |
| CNP | 3.47% (97.17%) | 0.08% | 0.01% | 0.01% | 0.10% |
| CS | 1.45% (82.22%) | 0.29% | 0.02% | 0.00% | 0.31% |
| MPN | 2.87% (99.88%) | 0.00% | 0.00% | 0.00% | 0.00% |
| NP | 27.74% (95.01%) | 1.42% | 0.03% | 0.00% | 1.45% |
| PP | 23.95% (97.63%) | 0.57% | 0.01% | 0.00% | 0.58% |
| S | 15.31% (76.53%) | 3.45% | 1.21% | 0.03% | 4.69% |
| VP | 5.84% (65.18%) | 2.59% | 0.50% | 0.02% | 3.11% |
| VZ | 1.17% (100.0%) | 0.00% | 0.00% | 0.00% | 0.00% |
| | $\Sigma^0 = 86.33\%$ | $\Sigma^1 = 9.17\%$ | $\Sigma^2 = 1.82\%$ | $\Sigma^3 = 0.06\%$ | $\Sigma^d = 11.05\%$ |

# Results: Continuous wordgroups in Dutch (CGN)



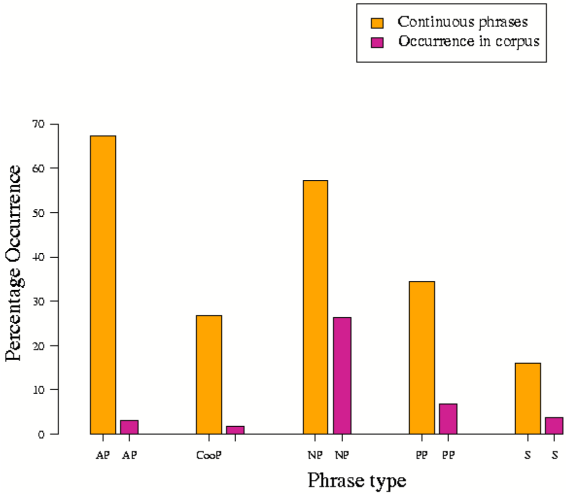
# Results: Discontinuous wordgroups in Dutch (CGN)



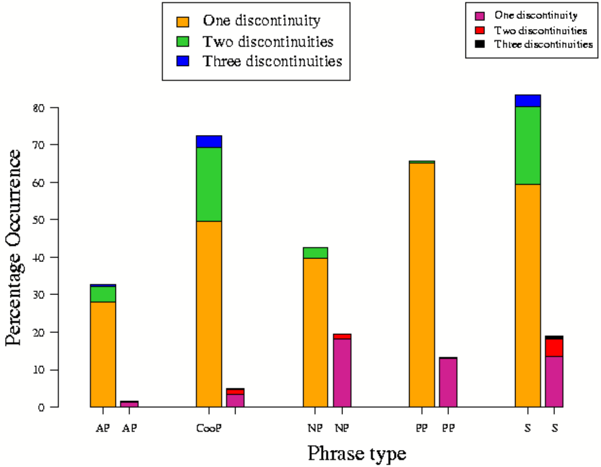
# Frequency (in %) of (dis)continuous occ. per total # occ. (CGN)

| **Type** | **0 holes** | **1 hole** | **2 holes** | **3 holes** | $\Sigma(discont)$ |
|---|---|---|---|---|---|
| ADVP | 2.05% (96.85%) | 0.07% | 0.00% | 0.00% | 0.07% |
| AP | 2.54% (89.07%) | 0.30% | 0.00% | 0.00% | 0.30% |
| CONJ | 2.37% (70.91%) | 0.85% | 0.11% | 0.01% | 0.97% |
| CP | 3.46% (94.92%) | 0.19% | 0.00% | 0.00% | 0.19% |
| DU | 11.33% (82.31%) | 2.21% | 0.21% | 0.02% | 2.44% |
| INF | 2.25% (40.96%) | 2.83% | 0.39% | 0.03% | 3.25% |
| MWU | 3.43% (98.06%) | 0.07% | 0.00% | 0.00% | 0.07% |
| NP | 17.63% (90.77%) | 1.71% | 0.08% | 0.00% | 1.79% |
| PP | 10.78% (86.60%) | 1.64% | 0.03% | 0.00% | 1.67% |
| PPART | 1.32% (60.27%) | 0.78% | 0.09% | 0.01% | 0.88% |
| REL | 0.97% (93.27%) | 0.07% | 0.00% | 0.00% | 0.07% |
| SMAIN | 12.46% (77.60%) | 3.08% | 0.46% | 0.05% | 3.59% |
| SSUB | 3.32% (64.98%) | 1.35% | 0.37% | 0.06% | 1.78% |
| SV1 | 3.70% (88.54%) | 0.42% | 0.06% | 0.00% | 0.48% |
| TI | 0.28% (24.33%) | 0.87% | 0.00% | 0.00% | 0.87% |
| | $\Sigma^0 = 77.89\%$ | $\Sigma^1 = 16.44\%$ | $\Sigma^2 = 1.80\%$ | $\Sigma^3 = 0.20\%$ | $\Sigma^d = 18.42\%$ |

# Results: Continuous wordgroups in Czech (PDT)



# Results: Discontinuous wordgroups in Czech (PDT)



# Frequency (in %) of (dis)continuous occ. per total # occ. (PDT)

| **Type** | **0 holes** | **1 hole** | **2 holes** | **3 holes** | $\Sigma(discont)$ |
|---|---:|---:|---:|---:|---:|
| AP | 3.10% (67.19%) | 1.29% | 0.19% | 0.03% | 1.51% |
| CooP | 1.82% (26.70%) | 3.38% | 1.34% | 0.22% | 4.94% |
| NP | 26.29% (57.29%) | 18.28% | 1.20% | 0.09% | 19.57% |
| PP | 6.90% (34.33%) | 13.10% | 0.11% | 0.00% | 13.21% |
| S | 3.64% (16.10%) | 13.41% | 4.71% | 0.72% | 18.84% |
| | $\Sigma^0 = 41.75\%$ | $\Sigma^1 = 49.46\%$ | $\Sigma^2 = 7.55\%$ | $\Sigma^3 = 1.06\%$ | $\Sigma^d = 58.07\%$ |

# Comparison

| **Language** | $\Sigma^0$ | $\Sigma^1$ | $\Sigma^2$ | $\Sigma^3$ |
|---|---:|---:|---:|---:|
| English | 96.38% | 1.02% | 0.01% | 0.00% |
| German | 86.33% | 9.17% | 1.82% | 0.06% |
| Dutch | 77.89% | 16.44% | 1.80% | 0.20% |
| Czech | 41.75% | 49.46% | 7.55% | 1.06% |

- **Confirmation of null hypothesis**
- **German, Dutch**
  - VZ is continuous (nested) $\Sigma^d$=0.00%
  - INF, TI are mostly 'discontinuous' (cross-serial) ($\Sigma^d$=4.12%)
- **Nominal, verbal groups**
  - Nominal groups mostly continuous in E, G & D (99.39%, 95.01%, 90.77%)
  - Verbal groups show increase in discontinuity with increase in flexibility
- **Prediction: CFG more and more inadequate**

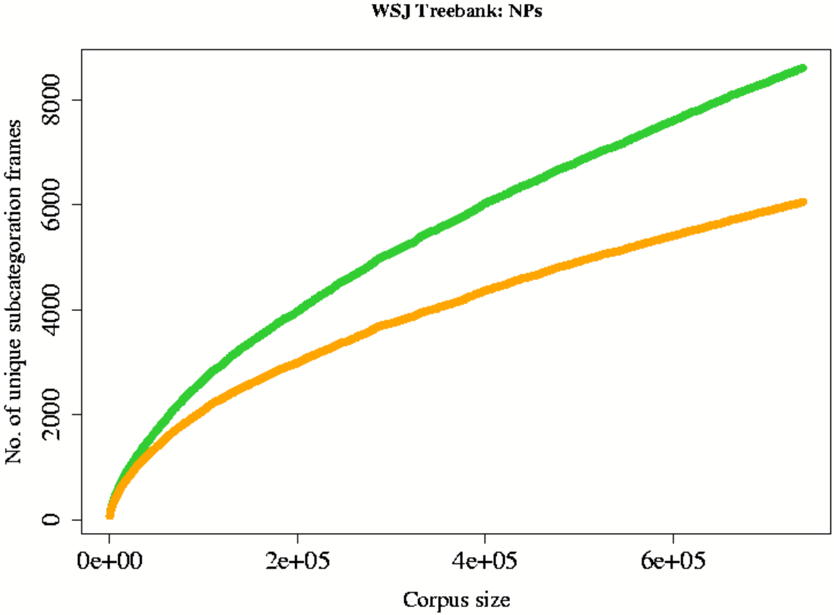
# Investigation: Scrambling

- **Null hypothesis: More freedom $\Rightarrow$ more scrambling**
- **Simple approach**
  - Ordered modification context: Dependency *tree*
  - Unordered modification context: Dependency *mobile*
  - variability factor $= log_2\left(\frac{|\text{unique ordered MCs}|}{|\text{unique unordered MCs}|}\right)$
  - Variability factor: 0 means no scrambling, >0 scrambling
- **Measurements**
  - Variability factor for MCs per head type
  - Growth of $|\text{ordered MCs}|$, $|\text{unordered MCs}|$

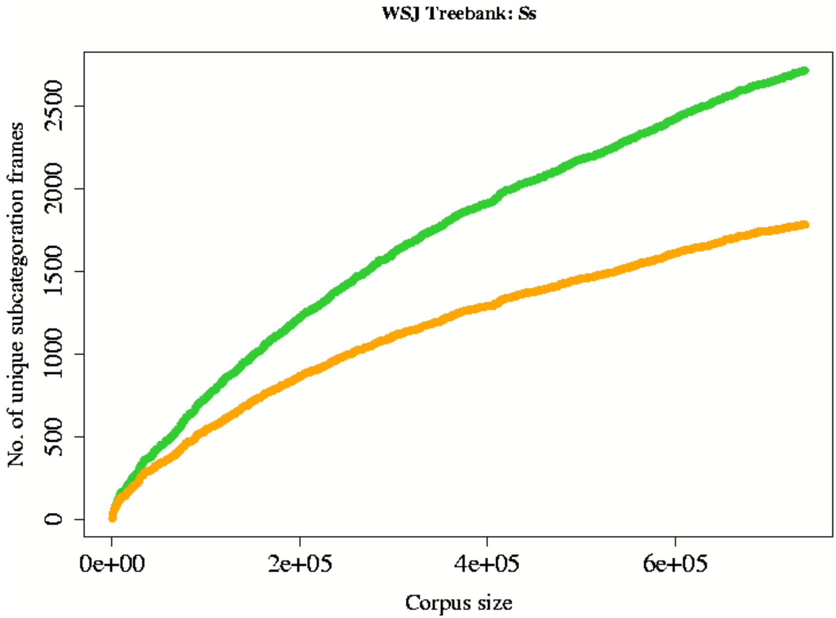
# Results: Variability factors for English (WSJ)

| **Type** | **Unique ordered** | **Unique unordered** | **Var. factor** |
|---|---|---|---|
| ADJP | 23 | 20 | 0.14 |
| NP | 409 | 303 | 0.30 |
| PP | 43 | 33 | 0.26 |
| QP | 95 | 70 | 0.31 |
| S | 445 | 325 | 0.31 |
| SBAR | 33 | 27 | 0.20 |
| VP | 561 | 452 | 0.22 |
| VPn | 99 | 76 | 0.26 |
| VPto | 2 | 1 | 0.69 |

# Results: Growth of unique NP mod.ctxts for English (WSJ)



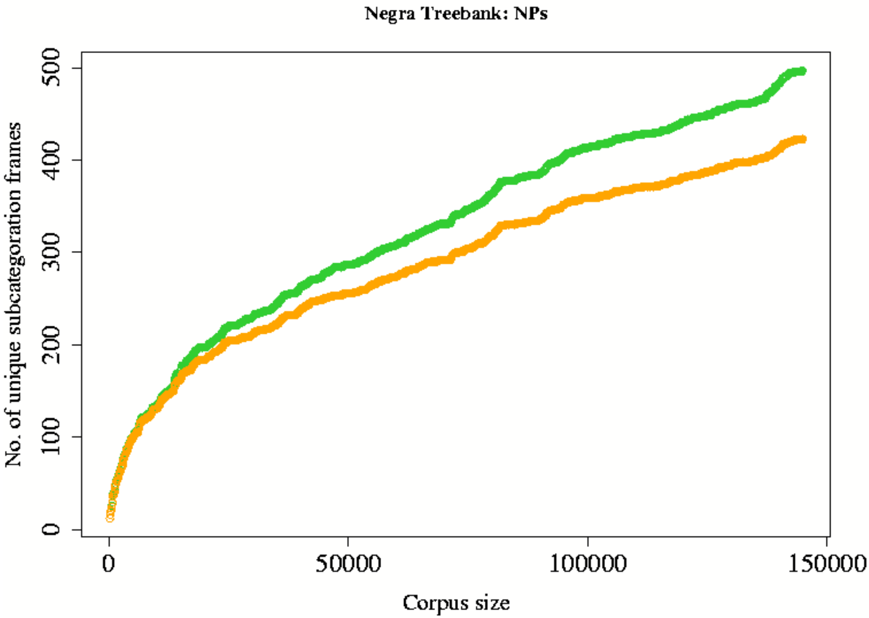
## Results: Growth of unique S mod.ctxts for English (WSJ)



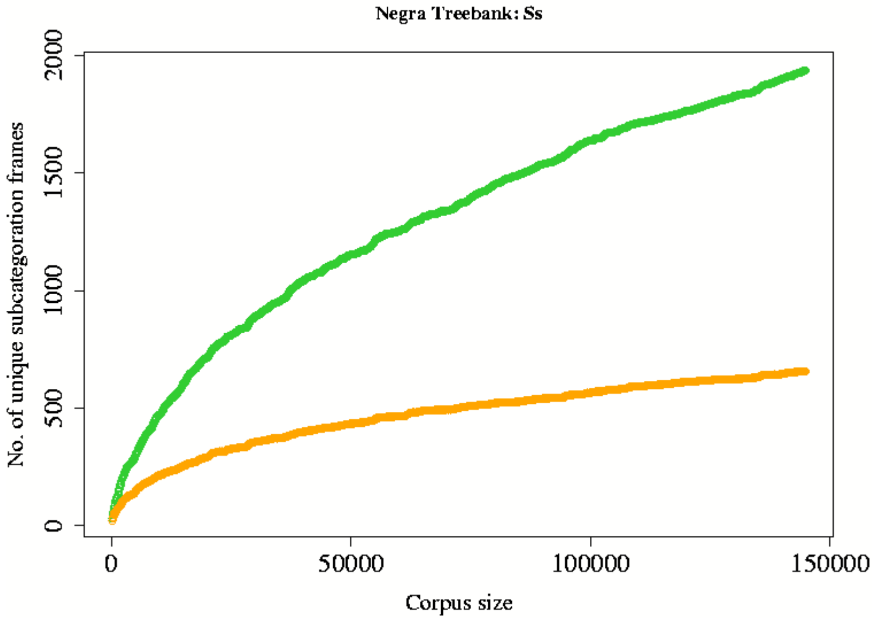
# Results: Variability factors for German (NEGRA)

| Type | Unique ordered | Unique unordered | Var. factor |
|---|---|---|---|
| AP | 77 | 52 | 0.39 |
| AVP | 9 | 8 | 0.12 |
| CNP | 21 | 17 | 0.21 |
| CS | 4 | 1 | 1.39 |
| MPN | 4 | 4 | 0.00 |
| NP | 229 | 193 | 0.17 |
| PP | 127 | 109 | 0.15 |
| S | 688 | 248 | 1.02 |
| VP | 211 | 76 | 1.02 |

# Results: Growth of unique NP mod.ctxts for German (NEGRA)



# Results: Growth of unique S mod.ctxts for German (NEGRA)

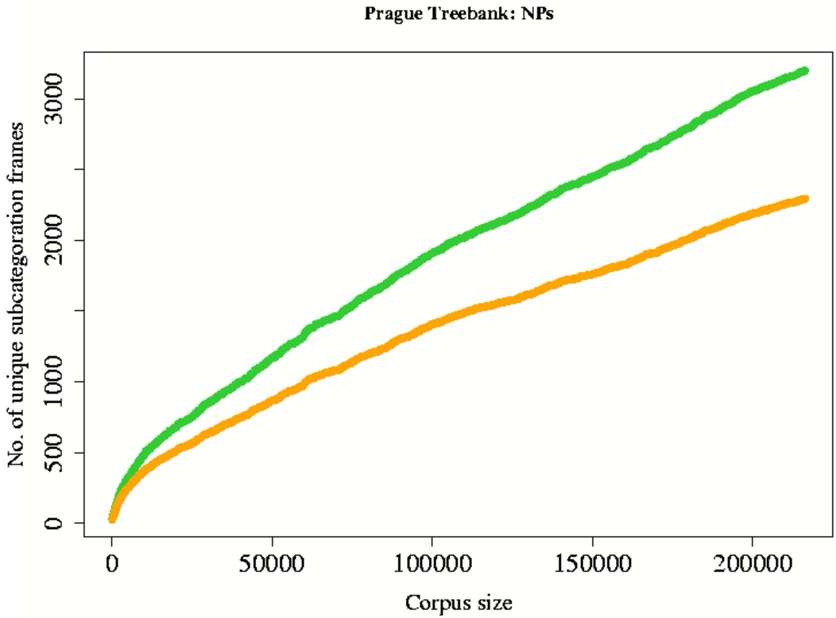


## Results: Variability factors for Czech (PDT)

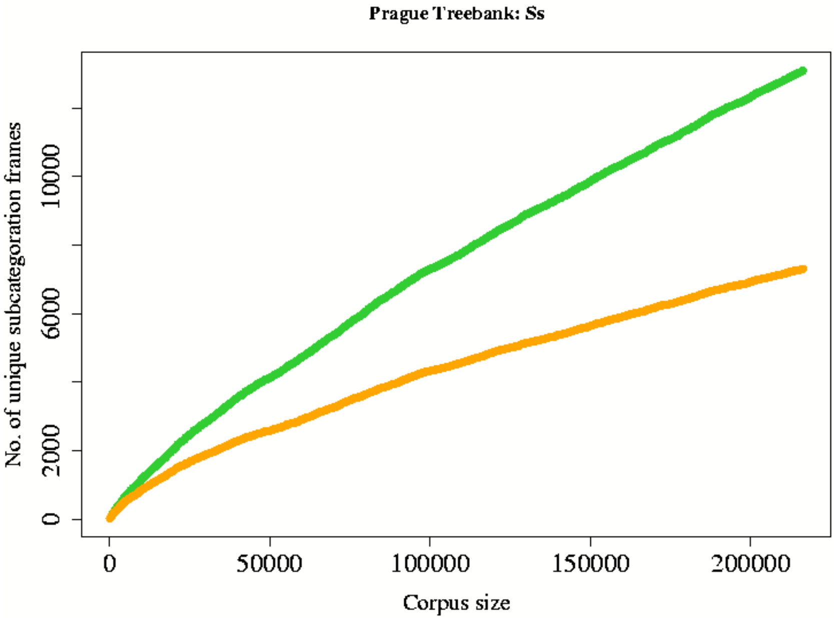
| Type | Unique ordered | Unique unordered | Var. factor |
|---|---|---|---|
| AP | 657 | 576 | 0.13 |
| CooP | 1973 | 1523 | 0.26 |
| NP | 2400 | 1686 | 0.35 |
| PP | 169 | 111 | 0.42 |
| S | 11350 | 6455 | 0.56 |

- **Surprising results**
  - Nominal groups behave similarly to English, German – expected
  - Verbal groups: English $\approx 2.5\times$ but German $\approx 0.5\times$
- **Possible reasons**
  - S class too indiscriminate
  - Genuine division of labor?

# Results: Growth of unique NP mod.ctxts for Czech (PDT)



# Results: Growth of unique S mod.ctxts for Czech (PDT)



- **Division of labor**
  - Long-distance dependencies (nonlocal)
  - Scrambling (local)
- **Compaction**
  - Square-root growth rate for TBG rule set with corpus size (Krotov et al., 1998)
  - Similar observations here
  - Krotov et al's compaction vs. variability in linearization

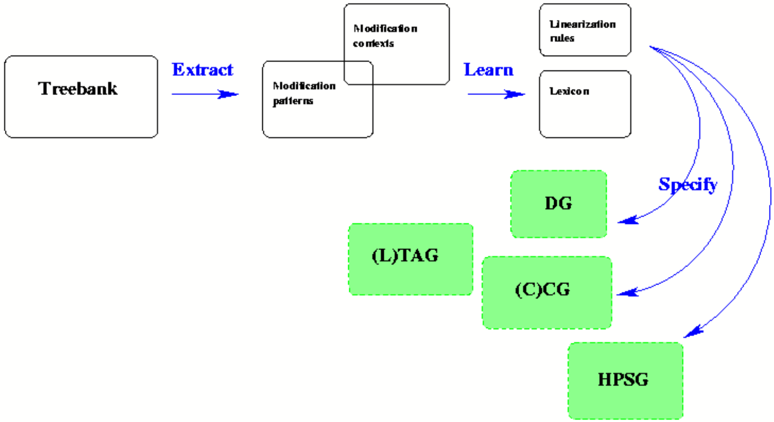
# Parsing results

- **"Treebank grammars"**
- **English**
  - Charniak (CFG), Xia *et al.* (LTAG), Hockenmaier, Clark *et al.* (CCG)
- **German**
  - Becker & Frank (topol.), Neumann (TAG/HPSG)
- **Dutch**
  - Adriaans (stat.), Moortgat & Moot (CTL)
- **Czech**
  - Sarkar & Zeman (stat.)
- **Lack of baseline information**
- **Applicability of formal algorithms**, e.g. Buszkowski & Penn

# Learning linearization rules

- **Data**
  - Discontinuous groups
  - Scrambling
- **Learning problem**
  - Given a set of trees (modification contexts,-patterns)
  - Restrictions ⇒ constancy of (partial) orderings
  - Variability ⇒ variation on canonical ordering
- **Approach**
  - Framework: 3-phase learning
  - Technique: multiple-sequence alignment (bioinformatics)

# 3-Phase Learning



# Robust learning of linearization rules

- **Variability in linearization**
  - Is a given linearization a variation of a canonical order?
- **Similar problem in bio-informatics**
  - Basic problem: *Pairwise alignment*
  - Given two sequences, possibly with gaps
  - Align the sequences
  - Related by chance or through evolution: Significance of scoring

- **Examples** (alignment to human alpha globin) (Durbin et al., 1998)

  *Clear similarity to human beta globin*

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

  *Structurally plausible alignment to leghaemoglobin from yellow lupin*

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

  *Spurious high-scoring alignment to a nematode glutathion S-transferase homologue*

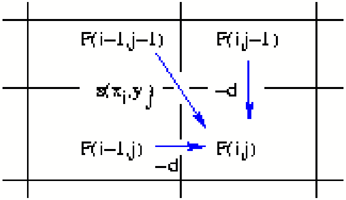
```
HBA_HUMAN GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
          GS+ + G +   +D L  ++ H+ D+  A +AL D     ++AH+
F11G11.2  GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

# Pairwise alignment (cont'd)

- **Scoring model**
  - Mutational processes: *substitution* and *insertion*, *deletion* (gaps)
  - Score: $\Sigma$ terms for each aligned pair of residues, plus terms for each gap
- **Examples**
  - Match model $M$, aligned pairs (a,b) have joint probability $p_{ab}$
  - "$a$ and $b$ derived independently from residue $c$ in ancestor"
  - Probability of alignment: $P(x, y|M) = \prod_i p_{x_i y_i}$
  - Gap penalties: linear $\gamma(g) = -gd$, affine $\gamma(g) = -d - (g-1)e$
- **Significance of scores**
  - Distinguish spurious from genuine alignments
  - Data-driven parametrization

# Pairwise alignment (cont'd)

- **Alignment algorithms**
  - Dynamic programming techniques
  - Find highest-scoring alignment of two sequences
  - Fill matrix F using solutions for optimal alignments of smaller subsequences




- **Different alignments**
  - Global alignment: Needleman-Wunsch algorithm
  - Local alignment: Smith-Waterman algorithm
  - Repeated matches (multiple local alignments), overlap matches

⇒ Hybrid match conditions

## Pairwise alignment and linearization

- **Pairwise alignment**
  - Similarity between two sequences (dependency trees)
  - Linguistically motivated gap penalty
- **Families and profiling**
  - Generalize alignment to form *families*
  - Create *profile* over family

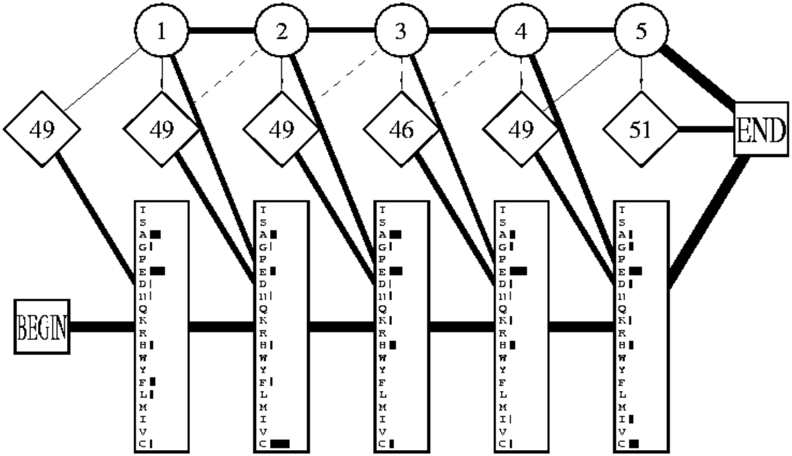
⇒ **Multiple sequence alignment**

- **Alignment of multiple sequences in colums**
  - Aligned substructures are 'homologous' – structurally similar
  - No single optimal alignment
  - Optimality depends on relatedness of training sequences ⇒ families
- **Scoring multiple alignment**
  - Position conservation
  - Sequences are not independent, but related
- **Multiple alignment by profile HMM training**

# Profile Hidden Markov Models

- ***Match***, ***insert*** **and** ***delete*** **states**
- **Key idea**
  - Profile of multiple aligned sequences
  - Transition, emission probabilities capture information about each position in alignment
- **Training**
  - Adaptations of standard model construction methods
  - Aligned or unaligned sequences as input
- **Profile useful for multiple sequence alignment**

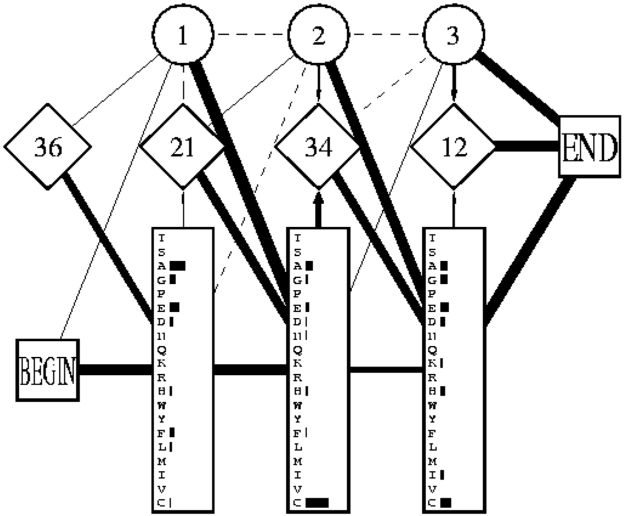
- **Family**
  - Sequences of equal length
  - Sequences of equal type
  - Variations (dep.trees) on dependency mobile
- **Tools**
  - Sequence Alignment Module (SAM) (Hughey and Krogh, 1996)
  - Translation edge-labels → single-characters
  - Treebank Perl scipts and Java classes




| | SB | HD | MO | CP | OC | OA |
|---|---|---|---|---|---|---|
| **Mapping table** | A | C | E | F | G | H |

# Aligned sequences of equal length: S, length 5

| | |
|---|---|
| sequence11 | EFADC |
| sequence32 | HECAE |
| sequence187 | ECAED |
| sequence246 | FHAEC |
| sequence247 | ECNEA |
| sequence284 | EFAGC |
| sequence291 | ECAEE |
| sequence317 | ECNAI |
| sequence392 | ACEHI |
| : | |

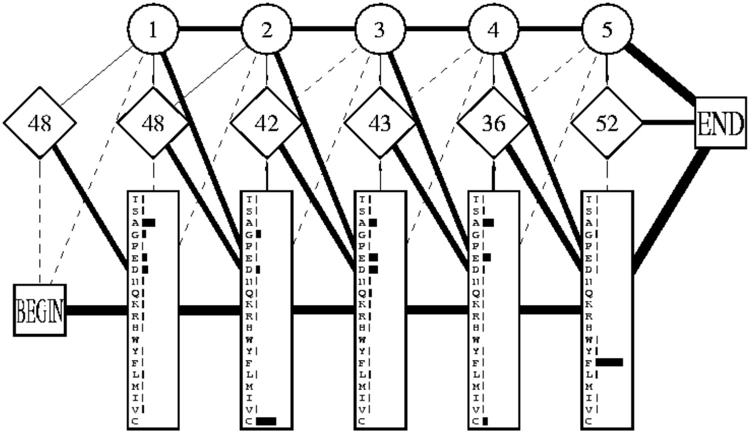



| | SB | HD | MO | CP | OC | OA |
|---|---|---|---|---|---|---|
| **Mapping table** | A | C | E | F | G | H |

# Aligned sequences of equal type: S

```
sequence38     A..C.......G.....
sequence39     A..C.......D.....
sequence46     A..C.......G.....
sequence48     G..C.......A.....
sequence56     A..C.......G.....
sequence58     -..C.......G.....
sequence62     A..C.......G.....
sequence66     A..D.......C.....
              :
```

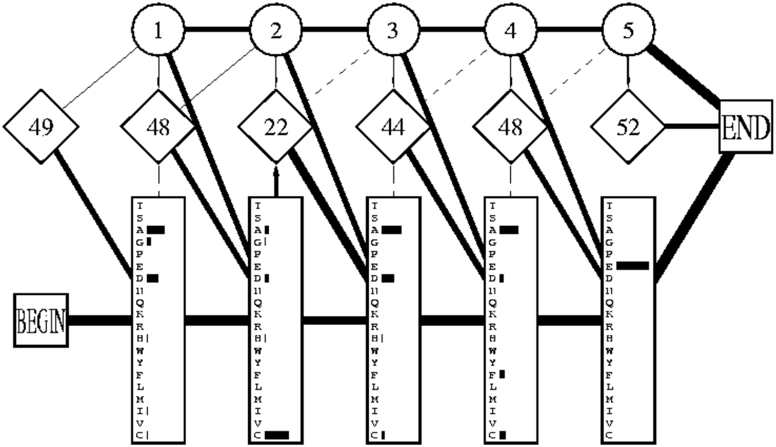
# Sequences of equal dependency mobile I (6 var)




`[HD:VMFIN, MO:ADV, MO:ADV, OC:VP, SB:NP]`

| | SB | HD | MO | CP | OC | OA |
|---|---|---|---|---|---|---|
| **Mapping table** | A | C | E | F | G | H |

# Sequences of equal dependency mobile II (7 var)

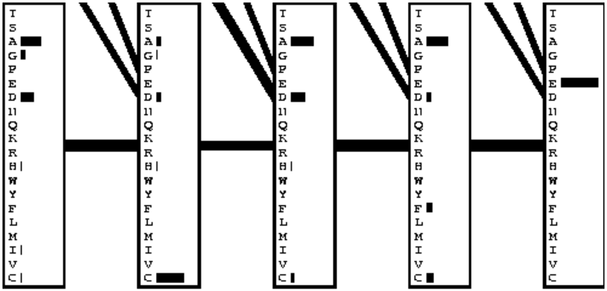



[HD:VVFIN, MO:PP, MO:PP, SB:NP]

| | SB | HD | MO | CP | OC | OA |
|---|---|---|---|---|---|---|
| **Mapping table** | A | C | E | F | G | H |

# Observations

- **From profiles to rules**
  - Soft rules: Use probabilities as preferences
  - Hard rules: discard patterns $\leq$ threshold
- **Example**
  - Hard rules yield classical topological model for II
  - {1 dep} HD {$n$ deps } { CP }
  - Canonical is SB-V order



# Conclusions & outlook

- **Potential approach to learning linearization rules**
  - Robust given scrambling, discontinuity
  - Linguistic intuitions
- **Improving data**
  - Phylogenetic trees $\overset{?}{\leftrightarrow}$ head alternation graphs
  - Sequence weighing (balancing data)
- **Improving computation**
  - Arbitrary sequence encoding
  - Linguistic data-based (affine) gapping scoring

# Contents

# References

Eugene Charniak. 1996. Tree-bank grammars. Technical report, Department of Computer Science, Brown University.

Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, England.

Mitchell Marcus et al. 1995. The Ultimate Penn Treebank Bible. Technical Report CD, Linguistic Data Consortium (UPenn). Technical Report.

Joseph H. Greenberg. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–114. The MIT Press, Cambridge, Massachusetts, second edition edition.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning*, pages 106–132. Karolinum, Prague, Czech Republic.

John A. Hawkins. 1983. *Word Order Universals*. Academic Press, New York, London, etc.

Richard Hughey and Anders Krogh. 1996. Hidden markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12:95–107.

Alexander Krotov, Mark Hepple, Robert Gaizauskas, and Yorick Wilks. 1998. Compacting the Penn treebank grammar. In *Proceedings COLING-ACL'98*, pages 699–703.

Geert-Jan M. Kruijff. 2001. *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, April.

Nelleke Oostdijk. 2000. The spoken dutch corpus. overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation Proceedings (LREC 2000)*, pages 887–894.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Applied Natural Language Processing 1997*, pages 88–95.

Susan Steele. 1978. Word order variation: A typological study. In Joseph H. Greenberg, editor, *Universals of Language. Volume 4: Syntax*, pages 585–624. Stanford University Press, Stanford, California.