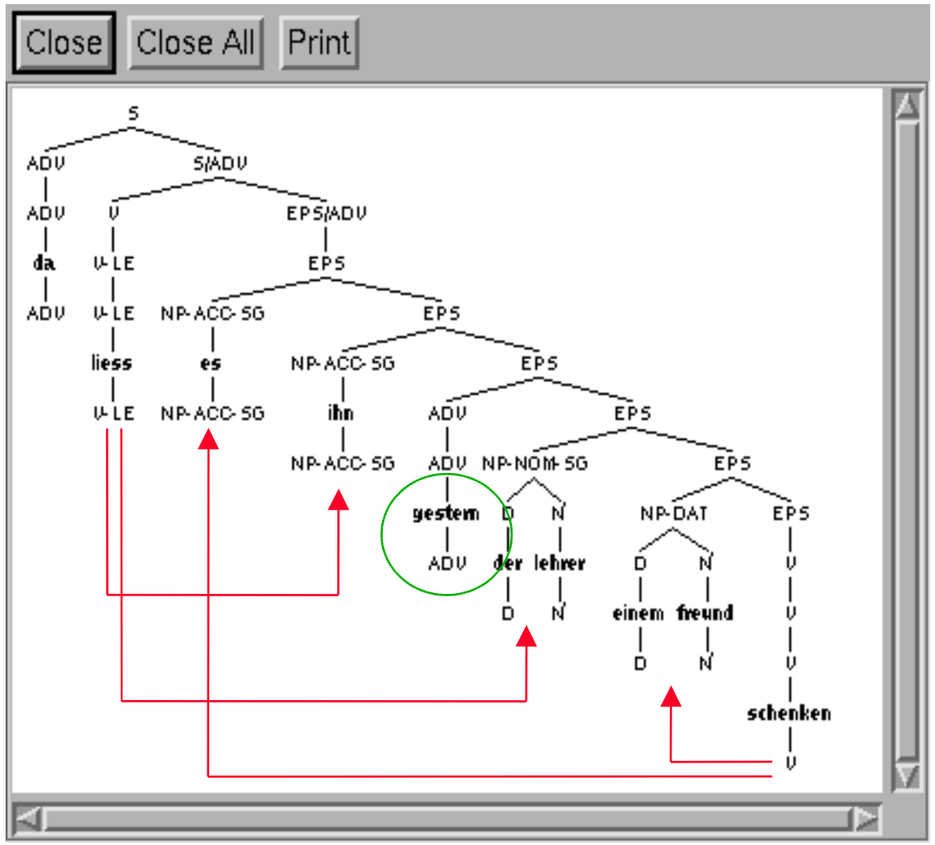
Efficient implementation of German verb placement in HPSG

Berthold Crysmann <crysmann@dfki.de>
Deutsches Forschungszentrum Künstliche Intelligenz (DFKI) GmbH

German – A nightmare for PS grammar

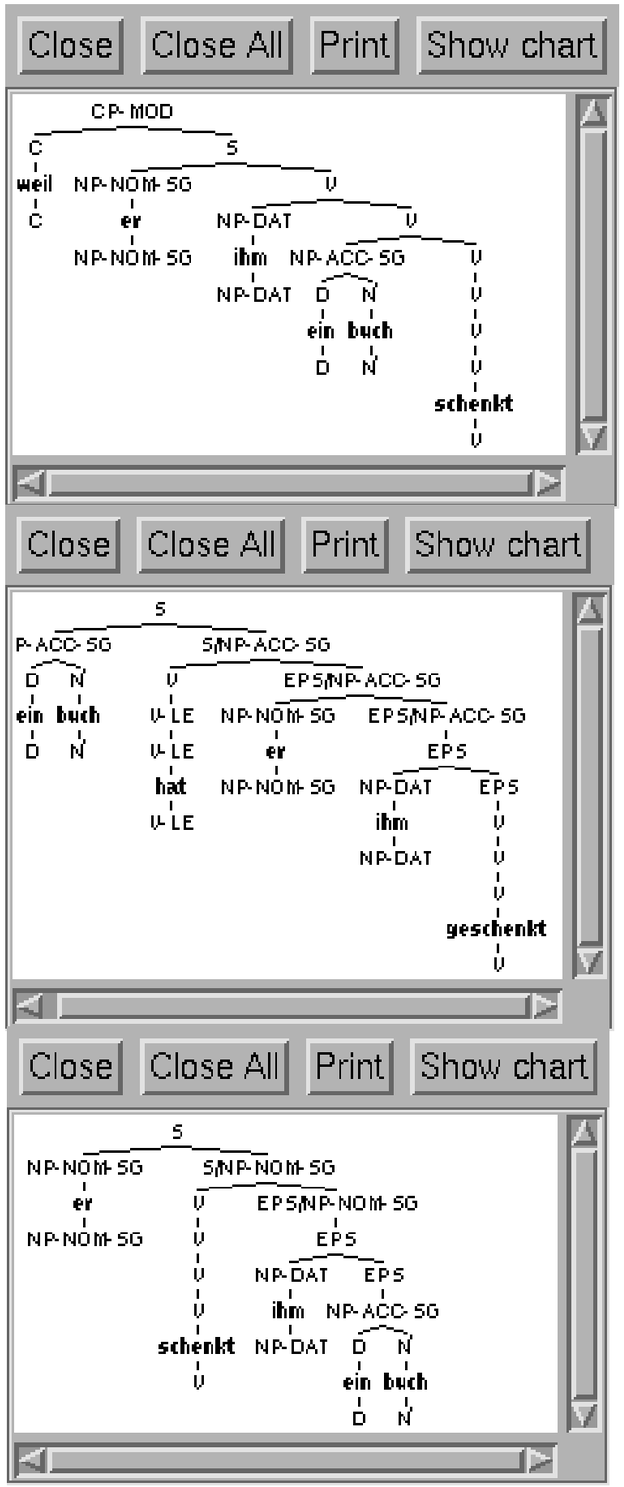
- **Variable verb placement**
- **Scrambling**
- **Clause union**
- **Discontinuous verb clusters**
- **Modifier interspersal**



Theoretical approaches

- **Order domains** (Reape 1994; Kathol 1995; Müller, 1999)
  - Currently no high-performance parsers available
  - Performance of existing Prolog implementations (Reape 1991, Müller 2002) cannot really compete with systems such as PET (factor 13 on VM)
  - Performance degrades w/o hard order constraints (Müller p.c.)
- **Flat structures** (Uszkoreit 1987, Nerbonne 1994)
  - Main linguistic motivation: free word order, no classical VP constituent
  - Problem: Interleaving of complements and modifiers
    - Statement of semantic composition clumsy (e.g., Kasper 1994)
    - Explosion of phrase structure rules (see Müller 2002)
- **Binary right-branching** (Kiss & Wesche 1991, Kiss 1995, Frank 1994, Netter 1998, Müller & Kasper 2000, Oliva 1992)
  - Permits recursive saturation of arguments
  - Modifiers and complements can be interspersed
  - Preferable for technical reasons
  - Discontinuous verb clusters related via simulated verb movement

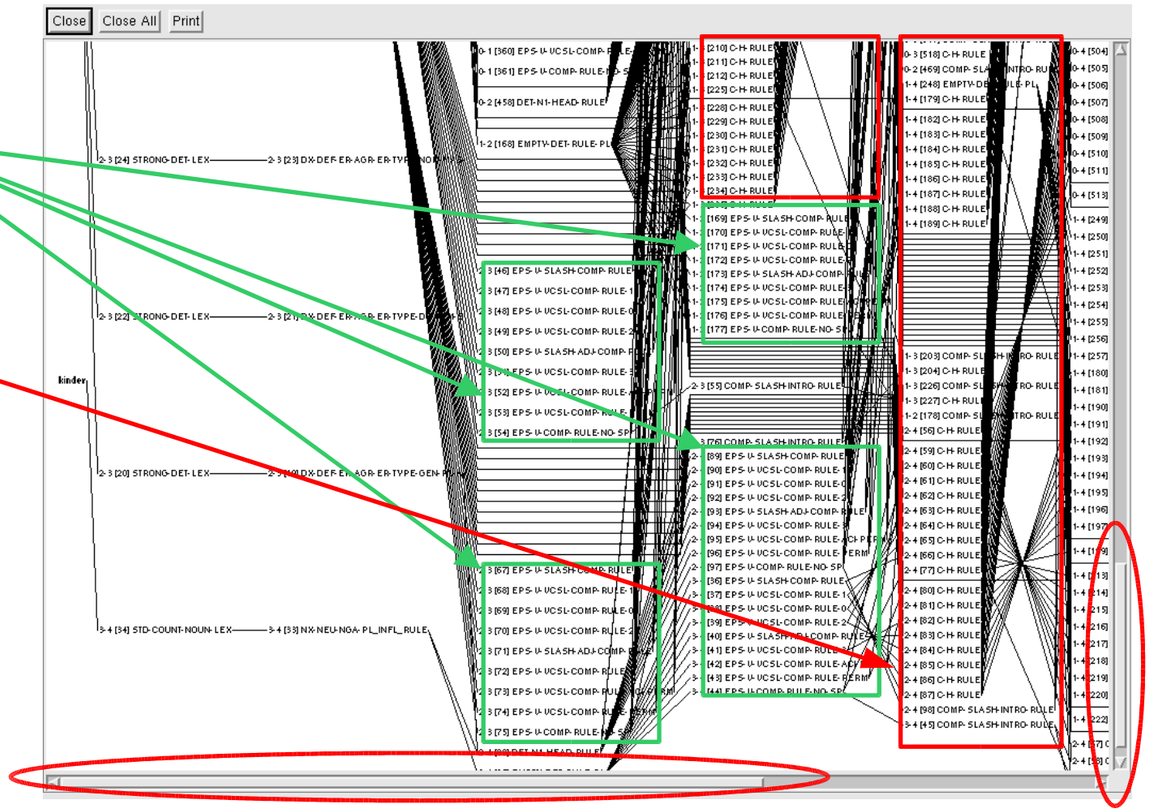
The original grammar (Müller & Kasper 2000)



- **Uniform right-branching analysis for**
  - Verb-final sentences
  - V1/V2 sentences w/ right sentence bracket
  - V1/V2 sentences w/o right sentence bracket
- **Initial verb related to final position by head-movement**
- **Main Problem: Efficiency**
- **Uniform head movement analysis requires**
  - Introduction of verb trace by means of unary rule (ϵ-rule)
  - Application to NPs/PPs/Adv for sentences w/o right bracket
- **Implementation as unary rules inelegant**
  - ϵ-rules must be differentiated according to the daughter's relation to the verb: total of 24 highly-tailored ϵ-rules

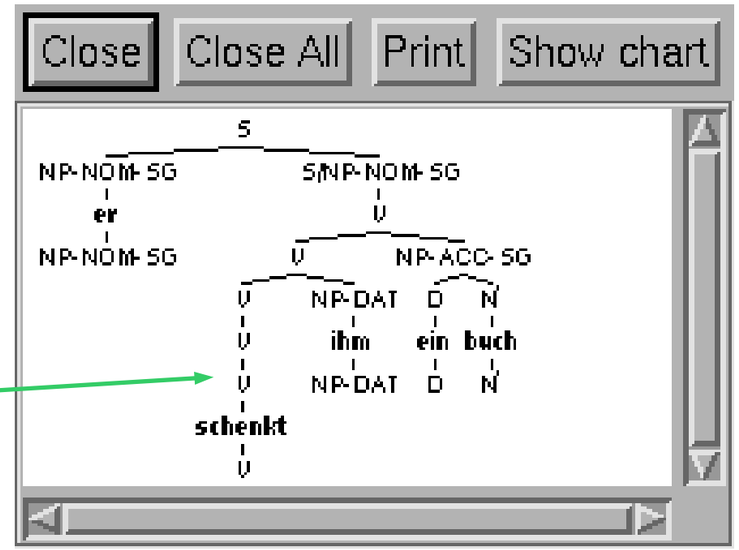
- **Application to non-verbal material**
  - Proliferation of ϵ-rules
  - Highly underspecified SUBCAT lists
  - ⇒ Combinatorial explosion
- **Ambiguity cannot be resolved locally**

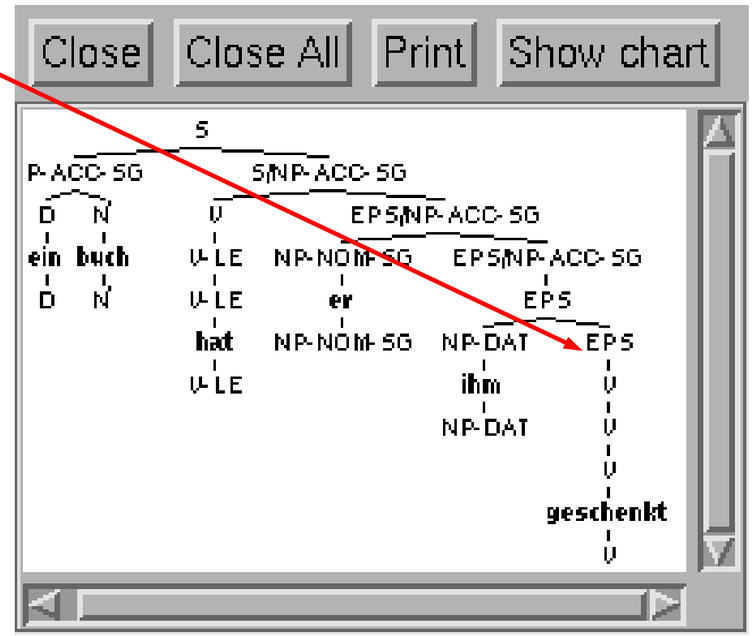
Partial chart for NP "der Lehrer der Kinder"

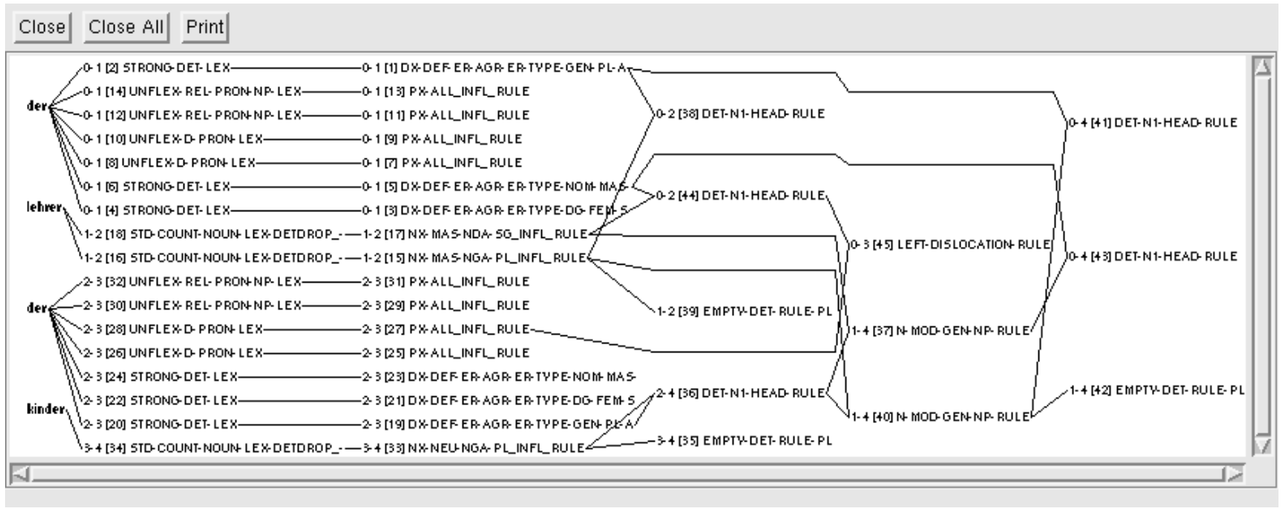


Revised grammar: Variable left and right branching

- **Design decision:**
  - Align position of the head with locus of highly restrictive information
  - Percolate less restricting information (eg. CONT)
- **Restrict head-movement analysis to constructions**
  - Where application of ϵ-rules can be restricted
  - Discontinuous cluster elements must be related
- **Asymmetric analysis**
  - Left branching for verb-initial sentences w/o right bracket
  - Right branching otherwise
  - ϵ-rules are restricted to apply to non-finite verbs and verb-particles
  - Subcat lists can be partially specified
- **Further optimisations:**
  - Separated SUBJ and COMPS:
    - Reduction in the number of permutation rules for cluster formation
  - Partial specification of SUBJ and COMPS list
    - At most one additional argument (subject) for final past participles
    - 1-2 additional arguments (1 subject, 1 complement) for infinitives
    - Full underspecification for stranded particles







Full parse chart for NP
der Lehrer der Kinder

Standard arguments for uniform right branching

- **Canonical order (Netter, 1998)**
  - Representation of canonical order as ordering on SUBCAT will require inversion by lexical rule
  - C-command relations are reversed
- **Modifier scope standardly determined left-to-right (e.g. Kasper, 1994, Müller, 2002)**

  Dass er ihn oft vergeblich gesucht hat

  `that he often looked for him in vain'

  Dass er ihn vergeblich oft gesucht hat

  `that it was vain that he looked for him often'
- **Scope argument relies crucially on the assumption that**
  - tree configurations, i.e., c-command relations, are the relevant representations
  - Left-to-right scope rule is not just a performance preference, however strong
  - Independent factors do not override tree-structural scopings (e.g. Focus)

Canonical order

- **Reversal of SUBCAT list easily implemented (5 rules)**
- **No order constraints in Müller & Kasper's grammar**
- **Unmarked order dependent on variety of factors**
  - Focus
  - Definiteness
  - Prosodic weight
- **Satisfactory treatment unlikely without soft constraints (Uszkoreit, 1987)**
  - Probably best achieved by means of corpus-derived probabilities for parse tree ranking

Modifier scope revisited

- **Left-to-right scope easily overridden, provided a plausible context**
  - Scopal over scopal

Da muß es schon erhebliche Probleme mit der Ausrüstung gegeben haben, da wegen schlechten Wetters ein Reinhold Messmer niemals aufgäbe.

'There must have been severe problems with the equipment, since someone like Reinhold Messmer would never give up just because of the bad weather.'

  - Scopal over intersective

Stefan ist wohl deshalb krank geworden, weil er äußerst hart wegen der Konferenz in Bremen gearbeitet hat.

'Stefan probably only became ill, because he worked extremely hard because of the upcoming conference in Bremen.'

- **Scope rule merely a strong preference**
- **Relegate preferred scope to performance component**
  - Grammar has relative modifier scope underspecified (see Müller & Kasper 2000)
  - Preferred scopings modelled via probabilities

Consequence:
Uniform right-branching empirically undermotivated

Experimental setup

- **Quantify efficiency gains by tests on NL corpora**
- **Regression tests on TSNLP test suite**
  - Version of the test suite with Müller's acceptability judgements
- **Performance experiment on Verbmobil data**
- **Planned:**
  - rerun earlier experiments on NEGRA newspaper corpus
  - Combine with ambiguity packing to study performance gains on increasing sentence length

Regression tests

- **Identical lexical ambiguity**
- **Comparable coverage**
- **Comparable overgeneration**

tsdb(1) `pet/de/tsnlp_sm/may_2002/03-08-02' vs. `pet/de/tsnlp_sm/apr_2003dis

Aggregate	(g)old				new			
	lexical Ø	parser Ø	in Ø	out Ø	lexical Ø	parser Ø	in Ø	out Ø
i-length in [15 .. 20]	3.31	1.00	38.1	0.0	3.31	1.00	38.1	0.0
i-length in [10 .. 15]	3.18	2.24	45.3	2.6	3.18	2.31	46.3	3.2
i-length in [5 .. 10]	3.69	2.21	69.3	17.3	3.69	2.14	69.9	17.8
i-length in [0 .. 5]	2.94	1.24	82.9	6.2	2.94	1.20	82.7	6.1
Total	3.26	1.66	73.8	7.3	3.26	1.62	74.0	7.3

(generated by [incr tsdb()] at 3-aug-03 (08:20))

Close LaTeX PostScript

Performance Experiment (vm-cd15)

- **Verbmobil corpus (spoken language dialogues) CD 15**
- **Native domain of Müller & Kasper (2000) grammar**
- **Moderate complexity:**
  - Maximum sentence length: 30-35 words
  - Average sentence length: 5.18 words (w/o punctuation)
- **Controlled variables**
  - Same machine (2.2 GHz Pentium 4, 1 GB RAM)
  - Lexical coverage
  - Quickcheck (computed individually on same corpus from unrelated domain)
- **Test parameters**
  - Upper limit of 70000 passive edges
    - PET parser (Callmeier 2000) delivers partial results when boundary is reached
  - Test results recorded with [incr tsdb()]
    - Records whether search space was exhaustively explored

Results (vm-cd15) – Coverage

- Identical lexical ambiguity
- Improved coverage
- Increased average sentence length (exhaustive)
  - Gold:3.9
  - New: 4.9
- Identical overgeneration on exhaustively parse items

tsdb(1) `pet/de/vm-cd15/may_2002/03-08-01' vs. `pet/de/vm-cd15/apr_2003dist

Aggregate	(g)old				new			
	lexical Ø	parser Ø	in Ø	out Ø	lexical Ø	parser Ø	in Ø	out Ø
i-length in [30 .. 35]	3.45	0.00	0.0	0.0	3.45	432.00	33.3	0.0
i-length in [25 .. 30]	3.44	0.00	0.0	0.0	3.44	192.00	12.5	0.0
i-length in [20 .. 25]	3.09	0.00	0.0	0.0	3.09	64.00	9.1	0.0
i-length in [15 .. 20]	3.10	12.75	6.7	0.0	3.10	50.65	48.3	12.5
i-length in [10 .. 15]	3.09	20.98	27.2	1.5	3.09	23.58	67.2	6.1
i-length in [5 .. 10]	3.18	6.56	76.7	10.0	3.18	6.70	82.8	11.7
i-length in [0 .. 5]	2.83	2.39	93.9	13.0	2.83	2.47	94.0	13.0
Total	**3.09**	**4.36**	**78.0**	**5.6**	**3.09**	**6.67**	**85.2**	**9.0**

(generated by [incr tsdb()] at 5-aug-03 (01:03))

Close LaTeX PostScript

Results (vm-cd15) – Performance

- Considerable performance gains: factor > 5

tsdb(1) `pet/de/vm-cd15/may_2002/03-08-01' vs. `pet/de/vm-cd15/apr_2003disbr/03-08-02' Per

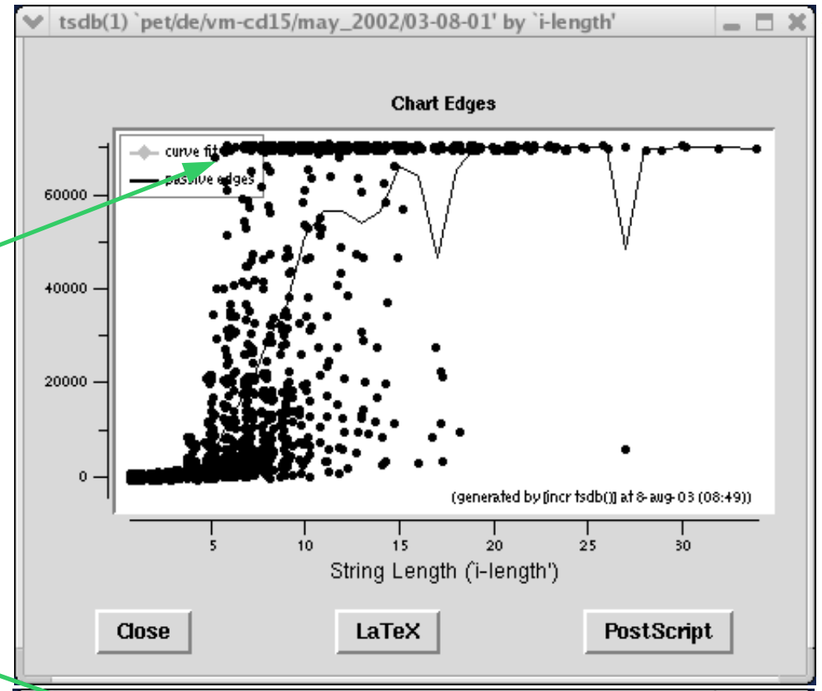
Aggregate	(g)old			new			reduction		
	tasks Ø	time Ø	space Ø	tasks Ø	time Ø	space Ø	tasks %	time %	space %
i-length in [30 .. 35]	197154	9.48	247487	168618	8.83	201198	14.5	6.8	18.7
i-length in [25 .. 30]	369728	18.85	243348	114950	6.40	171055	68.9	66.0	29.7
i-length in [20 .. 25]	271692	12.65	269190	101489	5.82	145805	62.6	54.0	45.8
i-length in [15 .. 20]	243884	13.09	227003	58767	3.21	83527	75.9	75.5	63.2
i-length in [10 .. 15]	164212	8.40	199524	27714	1.43	39913	83.1	83.0	80.0
i-length in [5 .. 10]	44659	2.23	63470	3583	0.18	4861	92.0	92.0	92.3
i-length in [0 .. 5]	612	0.03	762	101	0.00	139	83.5	81.6	81.8
Total	**47512**	**2.41**	**56703**	**8548**	**0.45**	**12116**	**82.0**	**81.1**	**78.6**

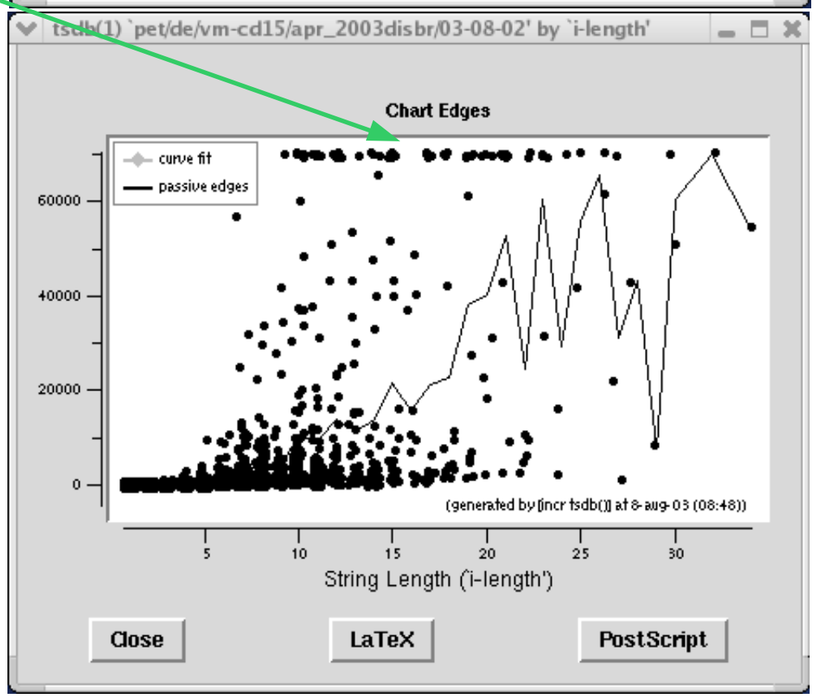
(generated by [incr tsdb()] at 5-aug-03 (01:05))

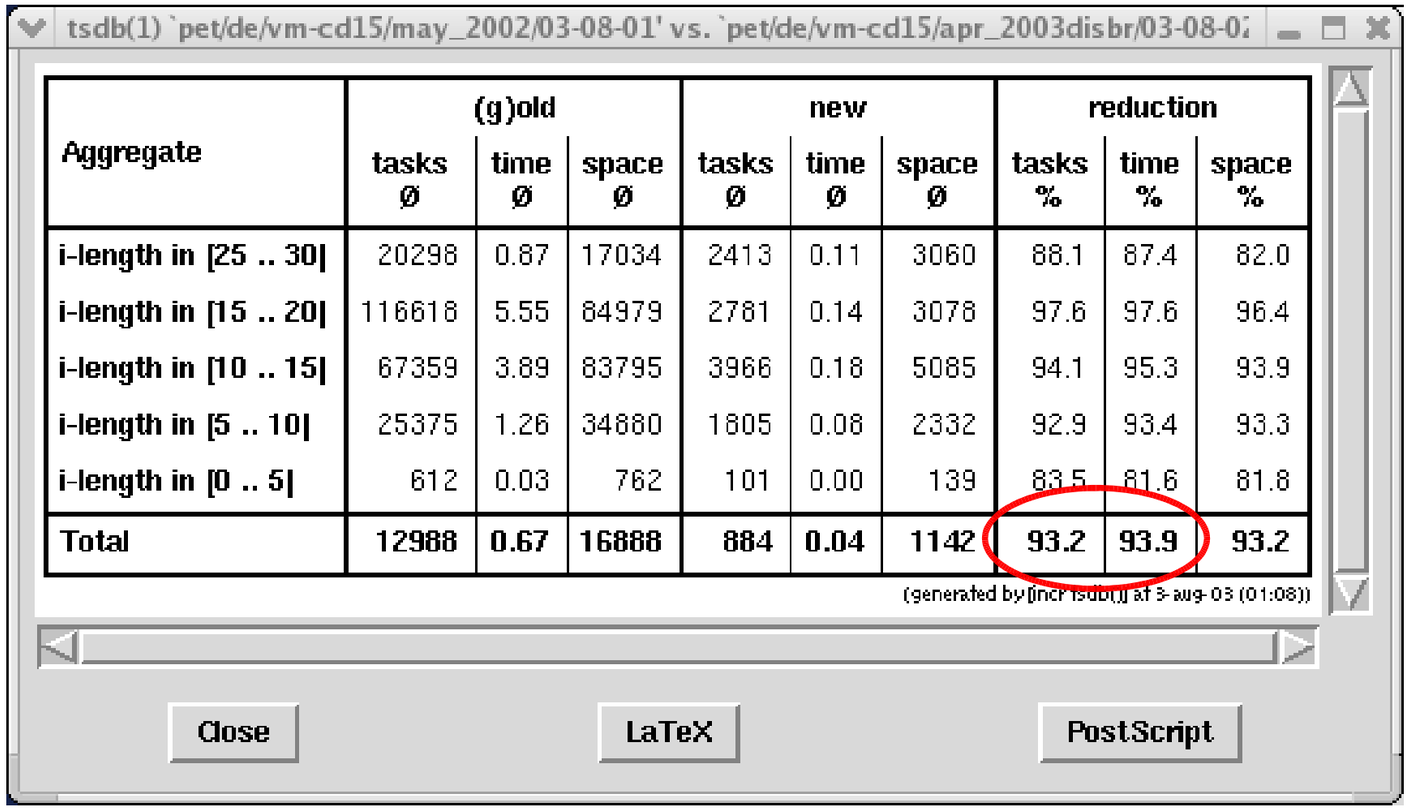
Close LaTeX PostScript

Results (vm-cd15) – Performance (intersection)

- Gold grammar benefits from ceiling effect (70,000 passive edges)
- Parse space not fully explored:
  - Original grammar (Gold): 16.1%
  - Revised grammar (New): 2.5%
- Intersection of exhaustively parsed sentences (<70,000 pedges)
- Elimination of ceiling effect
- True performance gain factor > 14.5







Aggregate	(g)old			new			reduction		
	tasks Ø	time Ø	space Ø	tasks Ø	time Ø	space Ø	tasks %	time %	space %
i-length in [25 .. 30]	20298	0.87	17034	2413	0.11	3060	88.1	87.4	82.0
i-length in [15 .. 20]	116618	5.55	84979	2781	0.14	3078	97.6	97.6	96.4
i-length in [10 .. 15]	67359	3.89	83795	3966	0.18	5085	94.1	95.3	93.9
i-length in [5 .. 10]	25375	1.26	34880	1805	0.08	2332	92.9	93.4	93.3
i-length in [0 .. 5]	612	0.03	762	101	0.00	139	83.5	81.6	81.8
Total	**12988**	**0.67**	**16888**	**884**	**0.04**	**1142**	**93.2**	**93.9**	**93.2**

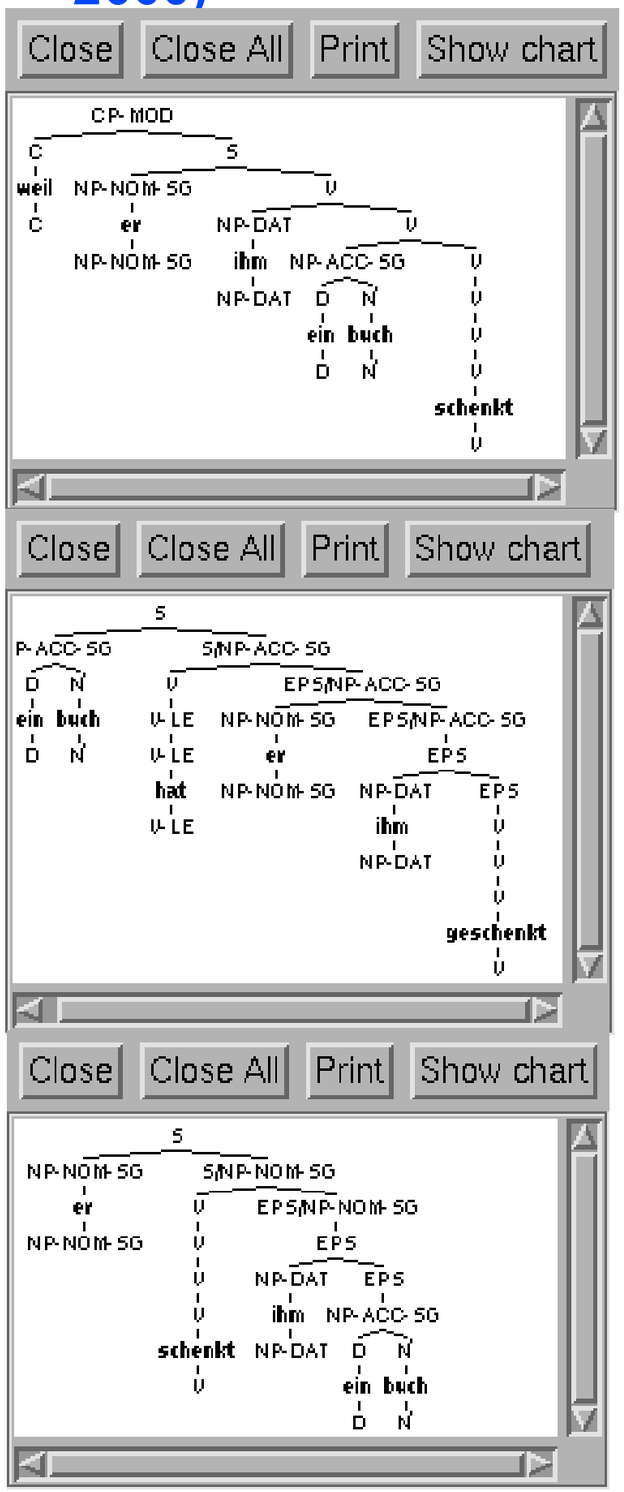
Experiment – Summary

- **Variable branching approach leads to significant performance and coverage gains over right-branching baseline**
  - Speed-up by more than factor 5 (with ceiling effects)
  - Roughly 15% increase in coverage on Verbmobil data
  - True performance gains considerably higher (> factor 14)
- **Combinatorial explosion induced by ε-rules defines a practical upper bound for original grammar (35 words)**
  - Increasing the number of passive edges to 100,000 leads to arithmetic overflow in timing function (vm-cd01): no reliable measurements
  - Coverage gains are minute (65% vs. 64% on vm-cd01)
- **Performance of revised grammar similar to LinGO ERG**
- **Speed-up factor compares well to recently reported advances**
  - Ambiguity packing (Oepen & Carroll 2000; Callmeier p.c.):
    factor 3-10 for ERG, depending on corpus
  - Deep-shallow integration (Frank, Becker, Crysmann, Kiefer & Schäfer 2003):
    factor 2.2 for first reading

Conclusion

- **Simple change in perspective towards verb placement leads to drastic performance increase**
- **Close genetic relationship of grammars makes it an ideal testing ground for alternative analyses**
- **High performance HPSG grammar engineering of utmost importance given renewed interest in deep processing**
- **Standard beliefs about right-branching German phrase structure turn out to be undermotivated**
  - LR-scoping, just like word order constraints, can be overridden
- **Performance gains should be regarded as empirical evidence in its own right**

The original grammar (Müller & Kasper 2000)



- **Uniform right-branching analysis for**
  - Verb-final sentences
  - V1/V2 sentences w/ right sentence bracket
  - V1/V2 sentences w/o right sentence bracket
- **Initial verb related to final position by head-movement**
- **Main Problem: Efficiency**
- **Uniform head movement analysis requires**
  - Introduction of verb trace by means of unary rule
  - Application to NPs/PPs/Adv for sentences w/o right bracket
- **Implementation as unary rules inelegant**
  - ϵ-rules must be differentiated according to the daughter's relation to the verb: total of 24 highly-tailored eps-rules

- **Application to non-verbal material**
  - leads to proliferation of ϵ-rules in the chart
  - Implies highly underspecified SUBCAT lists
  - Combinatorial explosion
- **Ambiguity cannot be resolved locally**

Parse chart for simple NP *der Lehrer*