How to Compare Treebanks

Sandra Kübler, Wolfgang Maier, Ines Rehbein
& Yannick Versley

LREC, May 2008

INDIANA UNIVERSITY
BLOOMINGTON

DCU
Dublin City University
Ollscoil Chathair Bhaile Átha Cliath

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Standardisation & Interoperability

- Creation of linguistic resources is extremely time-consuming
- Standardisation & interoperability

- One aspect of standardisation and interoperability
  - Adaptation of existing syntactic annotation schemes for new language ressources (e.g. Chinese Penn Treebank, Arabic Penn Treebank)
- But:
  - How to avoid importing flaws and weaknesses which might exist?
  - Are annotation schemes really universal?

We need to know more about syntactic annotation schemes and their impact on NLP applications

Standardisation & Interoperability

- Creation of linguistic resources is extremely time-consuming
- Standardisation & interoperability

- One aspect of standardisation and interoperability
  - Adaptation of existing syntactic annotation schemes for new language ressources (e.g. Chinese Penn Treebank, Arabic Penn Treebank)
- But:
  - How to avoid importing flaws and weaknesses which might exist?
  - Are annotation schemes really universal?

We need to know more about syntactic annotation schemes and their impact on NLP applications

Recent work

- Studies on the impact of treebank design on PCFG parsing:
  - *Kübler (2005), Maier (2006), Kübler et al. (2006)*
    Low PCFG parsing results (PARSEVAL) for the German NEGRA treebank imply that TüBa-D/Z is more adequate to support PCFG parsing
  - *Rehbein & van Genabith (2007)*
    Better PARSEVAL results for TüBa-D/Z reflect higher ratio of non-terminal/terminal nodes in the treebank

Results controversial, more extensive evaluation needed

Recent work

- Studies on the impact of treebank design on PCFG parsing:
  - *Kübler (2005), Maier (2006), Kübler et al. (2006)*
    Low PCFG parsing results (PARSEVAL) for the German NEGRA treebank imply that TüBa-D/Z is more adequate to support PCFG parsing
  - *Rehbein & van Genabith (2007)*
    Better PARSEVAL results for TüBa-D/Z reflect higher ratio of non-terminal/terminal nodes in the treebank

Results controversial, more extensive evaluation needed

Extensive evaluation

- of three different parsers
  - BitPar (Schmid, 2004)
  - LoPar (Schmid, 2000)
  - Stanford Parser (Klein & Manning, 2003)
- trained on two German treebanks
  - TiGer Release 2 (Brants et al., 2002)
  - TüBa-D/Z Release 3 (Telljohann et al., 2005)
- evaluated with
  - `evalb` (an implementation of PARSEVAL)
  - Leaf-Ancestor Metric (Sampson & Barbarczy, 2003)
  - Dependency-based Evaluation
  - Human evaluation

Outline

1. Data: TiGer & TüBa-D/Z
2. Experimental setup
3. Evaluation results
   - Constituent-based evaluation with PARSEVAL and LA
   - Dependency-based evaluation
   - Human evaluation

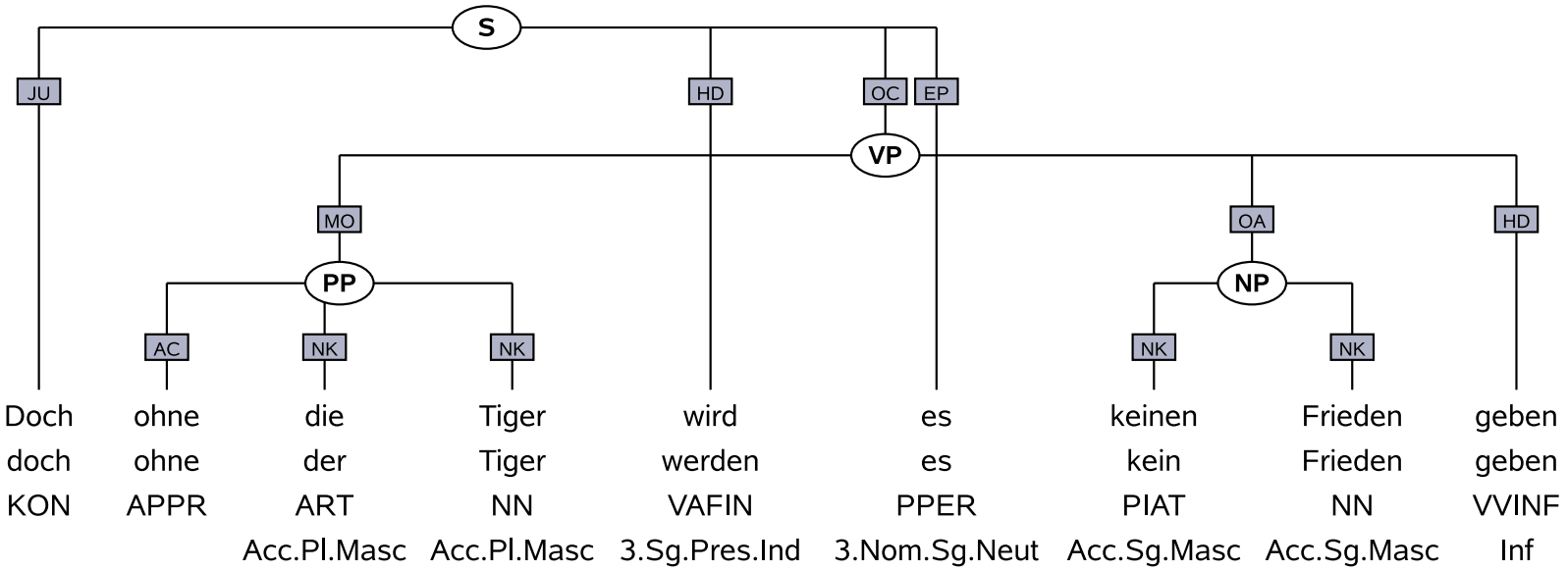
The Treebanks: TiGer and TüBa-D/Z

- Domain: German newspaper text
- POS tagset: STTS (Stuttgart-Tübingen Tag Set)

- Differences in annotation

	TiGer	**TüBa-D/Z**
Annotation:	flat	more hierarchical
LDD:	crossing branches	grammatical functions
Unary nodes:	no	yes
Topological fields:	no	yes

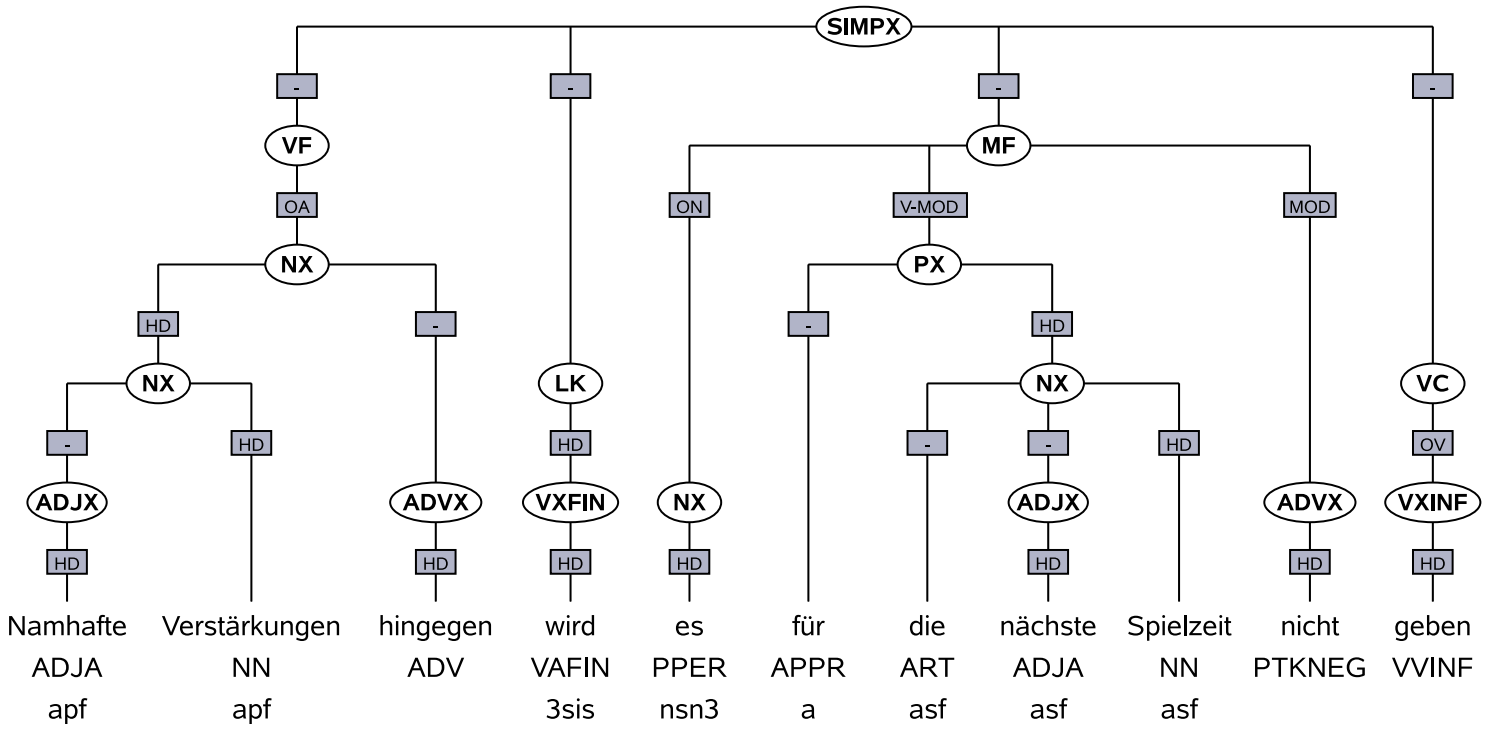
TiGer



But without the Tigers will it no peace give

"But without the Tigers there will be no peace."

TüBa-D/Z



Namable reinforcements however will it for the next playing season not give

"However, there won't be considerable reinforcements for the next playing time."

Experimental Setup

- Test Sets:
  - 2000 sentences from each treebank
- Training Sets:
  - 25 005 sentences from each treebank
- TiGer:
  - resolve crossing branches
  - insert preterminal nodes for all terminals with governable grammatical functions
- Train BitPar, LoPar and Stanford Parser on training sets
  - BitPar and LoPar: unlexicalised
  - Stanford: factored Model (PCFG+dependencies), hMarkov=1, vMarkov=2

Results for Constituent Evaluation

PARSEVAL and LA scores (2000 sentences)

	TiGer			TüBa-D/Z		
	Bit	**Lop**	**Stan**	**Bit**	**Lop**	**Stan**
evalb	74.0	75.2	77.3	83.4	84.6	88.5
LA	90.9	91.3	92.4	91.5	91.8	93.6

- `evalb` and LA: better results for TüBa-D/Z
- both measures show the same ranking:
  BitPar $<$ LoPar $<$ Stanford
- gap between LA results much smaller than between `evalb`

Discussion: PARSEVAL - LA

- PARSEVAL *(Black et al., 1991)*
  - divides number of matching brackets by overall number of brackets in the trees
  - more hierarchical annotation in TüBa-D/Z results in higher number of brackets
  - one mismatching bracket in TüBa-D/Z is punished less
- Leaf-Ancestor Metric *(Sampson & Barbarczy, 2003)*
  - string-based similarity measure based on Levenshtein distance
  - extracts path for each terminal node to the root node
  - computes the cost of transforming parser output paths into gold tree paths
  - edit cost is computed relative to path length → results in lower costs for same error for TüBa-D/Z

PARSEVAL and LA are biased towards TüBa-D/Z; Dependency evaluation should abstract away from particular encoding schemes

Discussion: PARSEVAL - LA

- PARSEVAL *(Black et al., 1991)*
  - divides number of matching brackets by overall number of brackets in the trees
  - more hierarchical annotation in TüBa-D/Z results in higher number of brackets
  - one mismatching bracket in TüBa-D/Z is punished less
- Leaf-Ancestor Metric *(Sampson & Barbarczy, 2003)*
  - string-based similarity measure based on Levenshtein distance
  - extracts path for each terminal node to the root node
  - computes the cost of transforming parser output paths into gold tree paths
  - edit cost is computed relative to path length → results in lower costs for same error for TüBa-D/Z

PARSEVAL and LA are biased towards TüBa-D/Z; Dependency evaluation should abstract away from particular encoding schemes

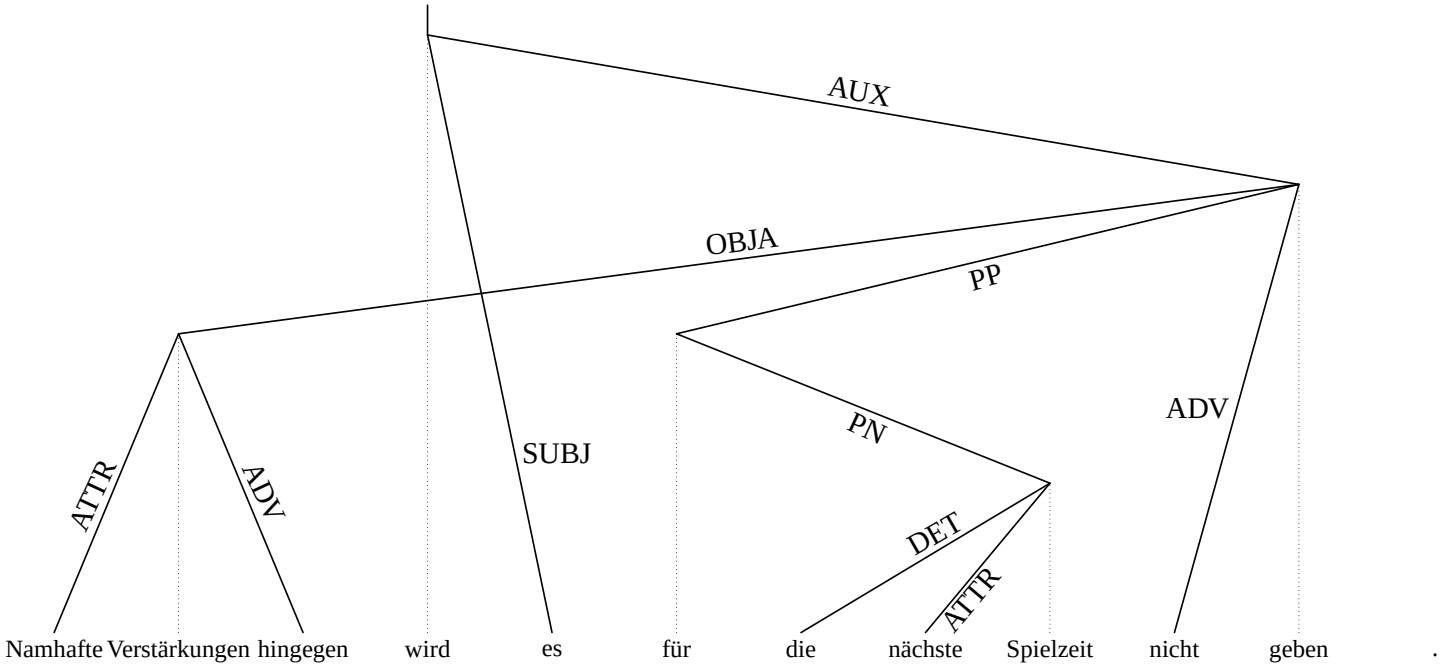
Discussion: PARSEVAL - LA

- PARSEVAL *(Black et al., 1991)*
  - divides number of matching brackets by overall number of brackets in the trees
  - more hierarchical annotation in TüBa-D/Z results in higher number of brackets
  - one mismatching bracket in TüBa-D/Z is punished less
- Leaf-Ancestor Metric *(Sampson & Barbarczy, 2003)*
  - string-based similarity measure based on Levenshtein distance
  - extracts path for each terminal node to the root node
  - computes the cost of transforming parser output paths into gold tree paths
  - edit cost is computed relative to path length → results in lower costs for same error for TüBa-D/Z

PARSEVAL and LA are biased towards TüBa-D/Z; Dependency evaluation should abstract away from particular encoding schemes

Dependency-Based Evaluation

- Original treebanks and parser output converted into dependencies
- 34 different dependency relations (Foth, 2003)
- Conversion with Depsy (Daum et al., 2004) and software by Versley (2005)



"However, there won't be considerable reinforcements for the next playing time"

Dependency-Based Evaluation: Results

PARSEVAL and LA scores (2000 sentences)

	TiGer			TüBa-D/Z		
	Bit	**Lop**	**Stan**	**Bit**	**Lop**	**Stan**
evalb	74.0	75.2	77.3	83.4	84.6	88.5
LA	90.9	91.3	92.4	91.5	91.8	93.6

Labeled/unlabeled dependency accuracy (2000 sentences)

	TiGer			TüBa-D/Z		
	Bit	**Lop**	**Stan**	**Bit**	**Lop**	**Stan**
Labelled Accuracy	78.8	80.5	81.6	71.3	72.8	75.9
Unlabelled Accuracy	83.0	84.5	85.6	81.7	83.4	86.8

Dependency-Based Evaluation: Results

- TüBa-D/Z gets slightly better results for unlabelled accuracy
- TiGer does better for labelled accuracy

- Results contradict constituent-based evaluation

- Human evaluation – How do the parsers perform on particular grammatical constructions?
  - Select sentences from both treebanks covering the same grammatical constructions
  - Evaluate how the parsers handle these particular constructions

Dependency-Based Evaluation: Results

- TüBa-D/Z gets slightly better results for unlabelled accuracy
- TiGer does better for labelled accuracy
- Results contradict constituent-based evaluation
- Human evaluation – How do the parsers perform on particular grammatical constructions?
  - Select sentences from both treebanks covering the same grammatical constructions
  - Evaluate how the parsers handle these particular constructions

TePaCoC - the Testsuite

- **Te**sting **Pa**rser Performance on **Co**mplex Grammatical **C**onstructions
  - Extraposed Relative Clauses (ERC)
  - Forward Conjunction Reduction (FCR)
  - Coordination of Unlike Constituents (CUC)
  - Noun PP Attachment (PPN)
  - Verb PP Attachment (PPV)
  - Subject Gap with Finite/Fronted Verbs (SGF)
- 200 sentences (100 from each treebank)
- The two annotation schemes make different design decisions to encode the same construction
  $\Rightarrow$ Criteria needed to evaluate grammatical constructions across treebanks

TePaCoC - Error Classification

- How to ensure inter-annotator agreement and reliability of human evaluation?
  ⇒ Error classification: describe categories for possible parser errors

Example: Extraposed Relative Clauses

Error description		TiGer	TüBa-D/Z
(A)	Clause not recognized as relative clause	Grammatical function incorrect	SIMPX label instead of R-SIMPX
(B)	Head noun incorrect	Attachment error	Grammatical function incorrect
(C)	Clause not recognized	Clause not recognized	Clause not recognized
(D)	Clause boundaries not correct	Span error	Span error

Results for Human Evaluation

	TiGer			TüBa-D/Z			
	Bit	**Lop**	**Stan**	**Bit**	**Lop**	**Stan**	**Freq.**
ERC	20	19	19	0	0	3	41
FCR	26	27	23	11	9	13	40
PPN	9	9	16	15	14	14	60
PPV	15	16	18	14	13	18	62
CUC	6	8	5	6	7	5	39
SGF	18	20	20	7	10	8	40

Table: Correctly parsed grammatical constructions in TiGer and TüBa-D/Z (human evaluation)

Conclusions

- Human evaluation correlates with dependency-based evaluation
- Human evaluation helps to trace error types back to underlying treebank design decisions
- Main findings:
  - TiGer benefits from the flat annotation which makes it more transparent for the parser (e.g. for ERC, FCR and SGF)
  - TüBa-D/Z suffers from the more hierarchical structure where relevant clues are embedded too deep in the tree
  - Additional layer of topological fields in TüBa-D/Z increases the number of possible attachment positions (and so possible errors)
  - Topological fields reduce number of rules in the grammar and improve the learnability especially for small training sets

Thank You!

Questions?

Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T.Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. *In Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306-311, 1991.

Boyd, Adriane. Discontinuity Revisited: An Improved Conversion to Context-Free Representations. *In Proceedings of the Linguistic Annotation Workshop (LAW 2007)* Prague, Czech Republic.

Brants, Sabine, and Silvia Hansen. 2002. Developments in the TiGer Annotation Scheme and their Realization in the Corpus. *In Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)* pp. 1643-1649 Las Palmas.

Briscoe, E. J., J. A. Carroll, and A. Copestake. 2002. Relational evaluation schemes. *In Proceedings Workshop 'Beyond Parseval - towards improved evaluation measures for parsing systems', 3rd International Conference on Language Resources and Evaluation*, pp. 4-38. Las Palmas, Canary Islands.

Carroll, J., E. Briscoe and A. Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. *In Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain. 447-454.

Daum, Michael and Kilian Foth and Wolfgang Menzel 2004. Automatic transformation of phrase treebanks to dependency trees. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004).* Lisbon, Portugal.

Dipper, S., T. Brants, W. Lezius, O. Plaehn, and G. Smith. 2001. The TiGer Treebank. *In Third Workshop on Linguistically Interpreted Corpora LINC-2001*, Leuven, Belgium.

Drach, Erich. 1937. *Grundgedanken der Deutschen Satzlehre.* Frankfurt/M.

Foth, K. 2003. Eine umfassende Dependenzgrammatik des Deutschen. Technical report, Fachbereich Informatik, Universität Hamburg.

, Klein, Dan and Chris Manning. 2003. Accurate Unlexicalized Parsing. *In Proceedings of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Kübler, Sandra. 2005. How Do Treebank Annotation Schemes Influence Parsing Results? Or How Not to Compare Apples And Oranges. *In Proceedings of FANLP 2005)*, Borovets, Bulgaria, September 2005.

Kübler, Sandra, Erhard Hinrichs, and Wolfgang Maier. 2006. Is it Really that Difficult to Parse German? *In Proceedings of the 2006 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2006), Sydney, Australia, July 2006.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, 10.707-10 (translation of Russian original published in 1965).

Maier, Wolfgang. 2006. Annotation Schemes and their Influence on Parsing Results. *In Proceedings of the COLING/ACL 2006 Student Research Workshop)*, Sydney, Australia.

Rehbein, Ines and Josef van Genabith. 2007. Treebank Annotation Schemes and Parser Evaluation for German. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* Prague, Czech Republic.

Sampson, Geoffrey, and Anna Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9 (4):365-380.

Schmid, Helmut. 2000. LoPar: Design and Implementation. *Arbeitspapiere des Sonderforschungsbereiches 340, No. 149*, IMS Stuttgart, July 2000.

Schmid, Helmut. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. *In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

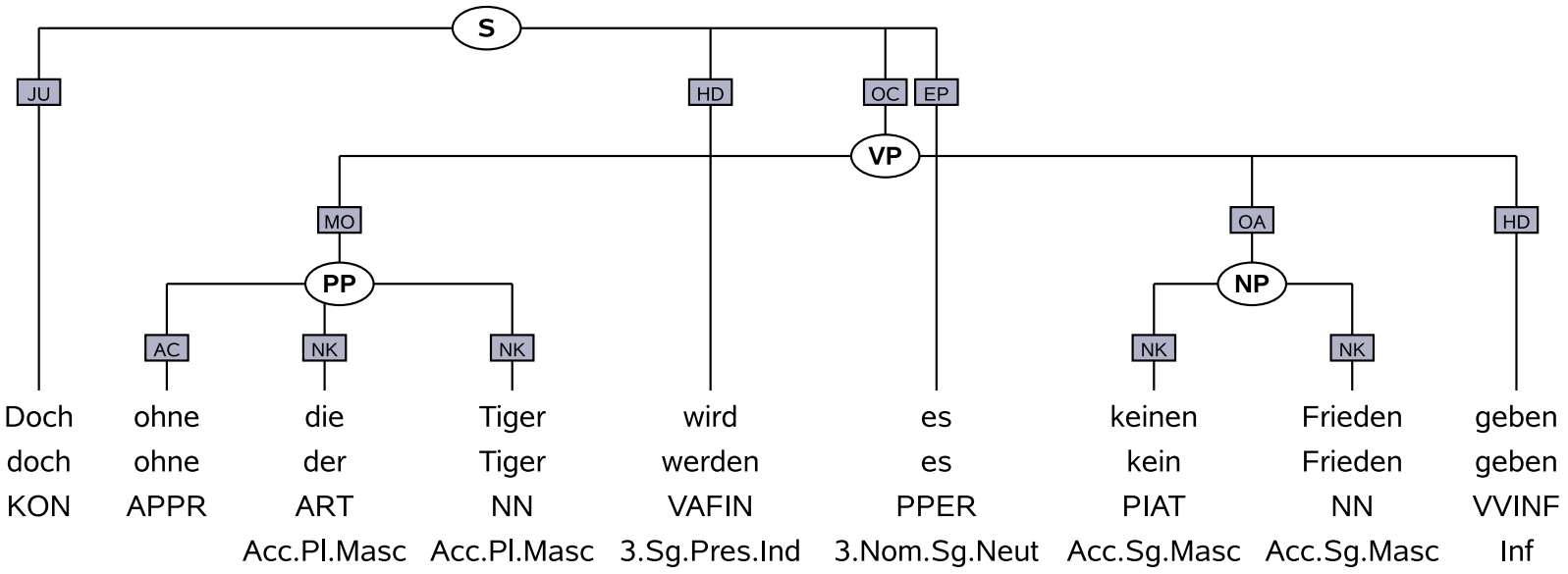
Skut, Wojciech, Brigitte Krann, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. *In Proceedings of ANLP 1997*, Washington, D.C.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2005. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Technical Report, IMS-CL, University Stuttgart, 1995.

Yannick Versley. 2005. Parser Evaluation across Text Types. *In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005).* Barcelona, Spain.

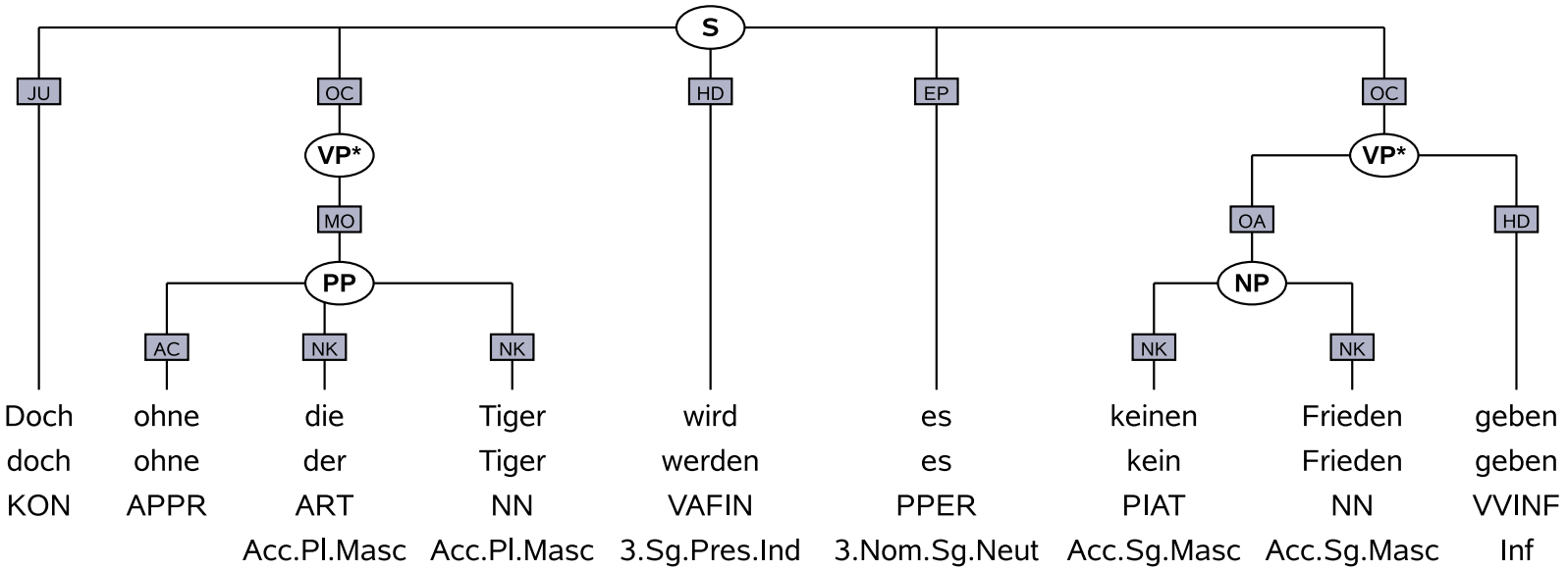
TiGer



But without the Tigers will it no peace give

"But without the Tigers there will be no peace."

TiGer



But without the Tigers will it no peace give

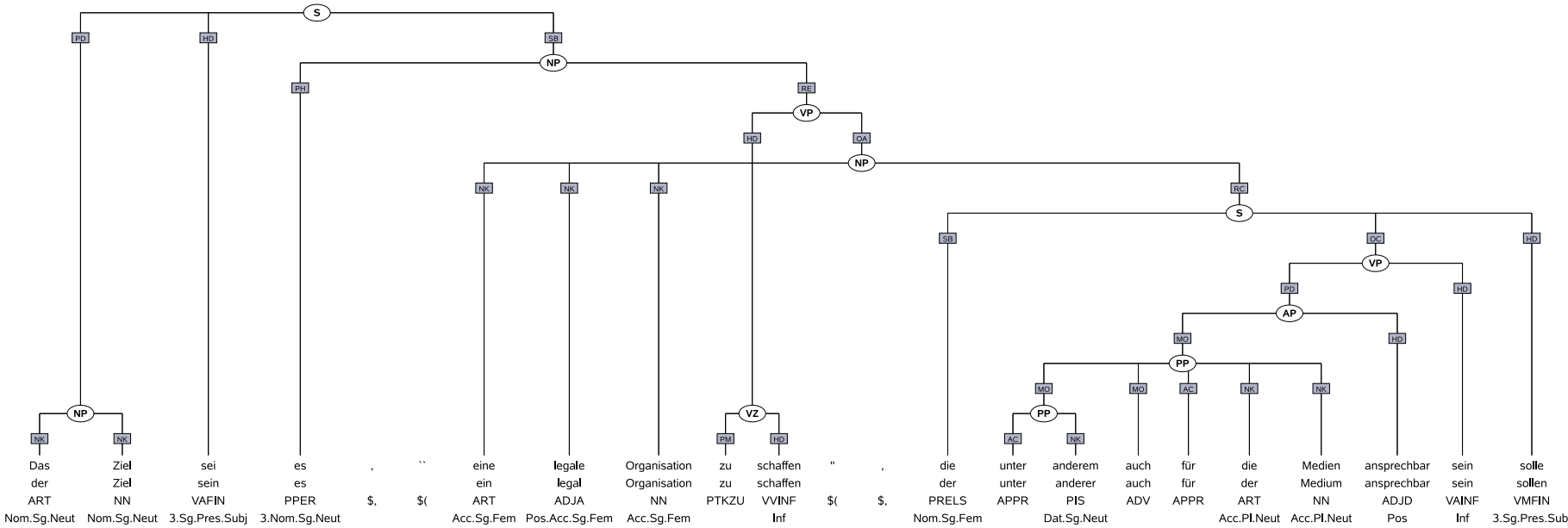
"But without the Tigers there will be no peace."

Dependency-Based Evaluation: Results

Dependency F-measure (2000 sentences):

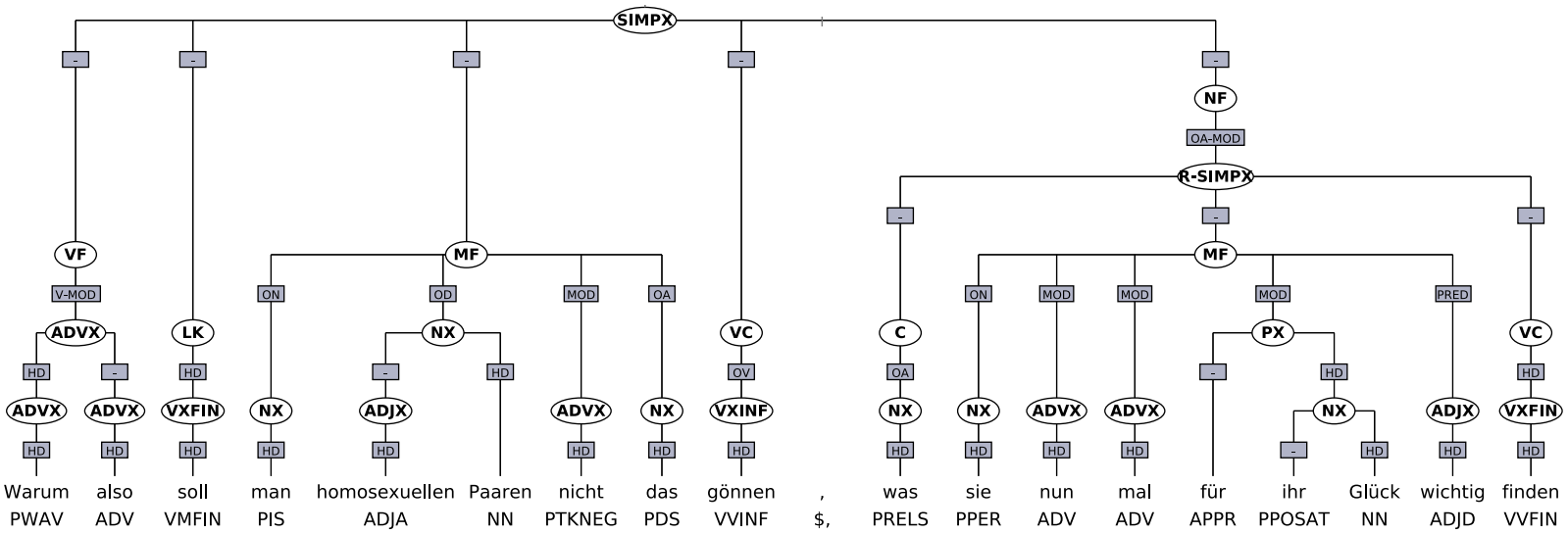
- nominal verb arguments (subjects and accusative/dative objects)
- PP attachment
- clause subordination (including infinitive and relative clauses as well as adjunct and argument subordinated clauses and argument full clauses)

	TiGer			**TüBa-D/Z**		
	Bit	**Lop**	**Stan**	**Bit**	**Lop**	**Stan**
SUBJ	80.2	81.1	78.7	74.6	75.3	76.1
OBJA	55.6	58.4	59.5	42.4	45.8	52.9
OBJD	11.6	11.5	14.1	12.9	13.3	13.1
PP	71.1	72.2	78.2	68.1	69.1	75.6
clause-sub.	57.0	58.2	60.9	45.8	47.5	52.1



(1) Das Ziel sei es, "eine legale **Organisation** zu schaffen", **die unter anderem auch**
The goal be it, "a legal organisation to create", which amongst others also
für die Medien ansprechbar sein soll.
for the media approachable be ought to

"The aim is to create a legal organisation which, amongst others, also ought to be approachable for the media."



(2) Warum also soll man homosexuellen Paaren nicht **das** gönnen, **was sie nun mal für ihr Glück wichtig finden?**
Why so shall one homosexual couples not that grant, which they now for their luck important find?

"So why shouldn't homosexual couples be granted what they think to be important to happiness."

Dependency-Based Evaluation for TePaCoC

		TiGer			TüBa-D/Z		
		Bit	Lop	Stan	Bit	Lop	Stan
LAS	ERC	76.2	76.0	77.4	71.6	71.8	71.1
	FCR	79.5	74.4	81.8	78.5	81.0	79.3
	PPN	76.8	79.7	87.0	75.5	76.1	76.1
	PPV	73.6	80.9	79.2	65.8	67.9	71.5
	CUC	65.2	67.0	70.7	57.5	63.0	60.9
	SGF	76.1	77.2	79.3	74.0	77.7	75.1
	ALL	**73.3**	**73.9**	**76.8**	**69.3**	**72.7**	**70.3**
UAS	ERC	81.1	80.8	82.0	79.1	80.5	79.1
	FCR	82.7	77.8	85.6	85.4	88.2	88.7
	PPN	84.2	86.4	89.3	84.8	85.3	85.9
	PPV	78.1	86.0	86.0	81.3	82.9	88.6
	CUC	69.7	71.5	74.7	66.1	72.0	73.6
	SGF	81.7	82.5	83.6	82.8	86.2	85.4
	ALL	**78.1**	**78.7**	**81.0**	**78.3**	**81.9**	**81.7**

Labeled/unlabeled dependency accuracy for the testsuite