UPPSALA
UNIVERSITET

Harvest Time

Explorations of the Swedish Treebank

Joakim Nivre

Uppsala University
Department of Linguistics and Philology

Thanks to Lars Ahrenberg, Evelina Andersson, Lars Borin, Elisabet Engdahl, Eva Forsbom, Sofia Gustafson-Čapková, Johan Hall, Janne Bondi Johannessen, Beáta Megyesi, Jens Nilsson, Filip Salomonsson, Anna Sågvall Hein, Reut Tsarfaty

A Personal TLT History

Sozopol, 2002 *What kinds of trees grow in Swedish soil?*

Växjö, 2003 *Theory-supporting treebanks*

Failed attempts to provide funding for a Swedish treebank ☹

Barcelona, 2005 *MaltParser: A language-independent system for data-driven dependency parsing*

More failed attempts to provide funding for a Swedish treebank ☹

Bergen, 2007 *Bootstrapping a Swedish treebank through cross-corpus harmonization and annotation projection*

Somewhat successful attempts to bootstrap a Swedish treebank ☺

Tartu, 2010 *Harvest time – what trees did in fact grow?*

Swedish Treebank 1.1

A low-budget treebank based on recycling:

- Talbanken
- The Stockholm-Umeå Corpus (SUC)

Two types of syntactic annotation:

- Phrase structure and grammatical functions
- Dependency structure

Availability:

- Free for research and education
- License required for SUC data
- Distributed by the Swedish Language Bank (http://spraakbanken.gu.se/eng/stb)

Outline of the Talk

The treebank:

- The raw material: Talbanken and SUC
- The recycling process
- The end result: Swedish Treebank

Explorations:

- Experiments in data-driven parsing
- Cross-framework parser evaluation

UPPSALA
UNIVERSITET

Swedish Treebank

The Swedish Treebank Project

Treebanking by recycling existing corpora:

- Talbanken – largest treebank (100k tokens)
- SUC – largest annotated corpus (1.2M tokens)
- Merge, harmonize and project missing annotation

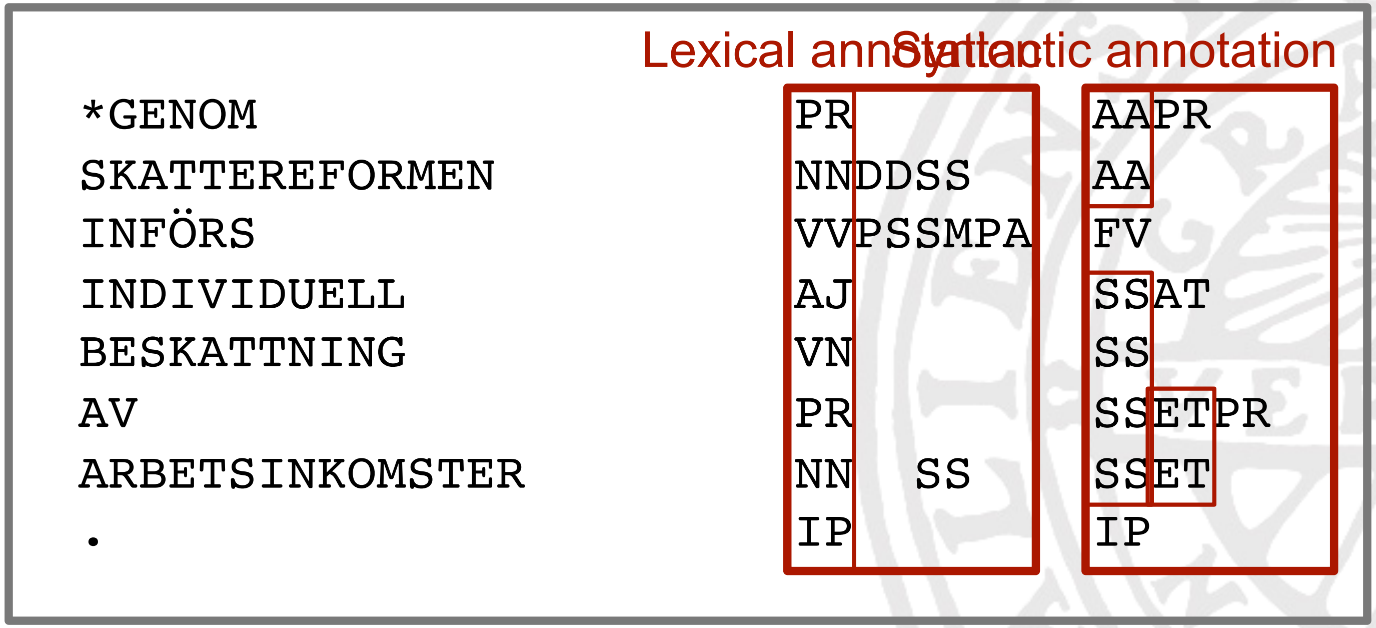
Collaboration between two projects:

- Methods and Tools for Grammar Extraction (Uppsala University)
- Inductive Dependency Parsing (Växjö University)

UPPSALA
UNIVERSITET

Talbanken

- Team led by Ulf Teleman, Lund University, 1970s
- Written and spoken Swedish (350k tokens)
  - Professional prose section (100k tokens)
- Annotation according to MAMBA [Teleman 1974]:
  - Lexical: parts of speech (PoS) + morphosyntactic features (MSF)
  - Syntactic: grammatical functions (GF)



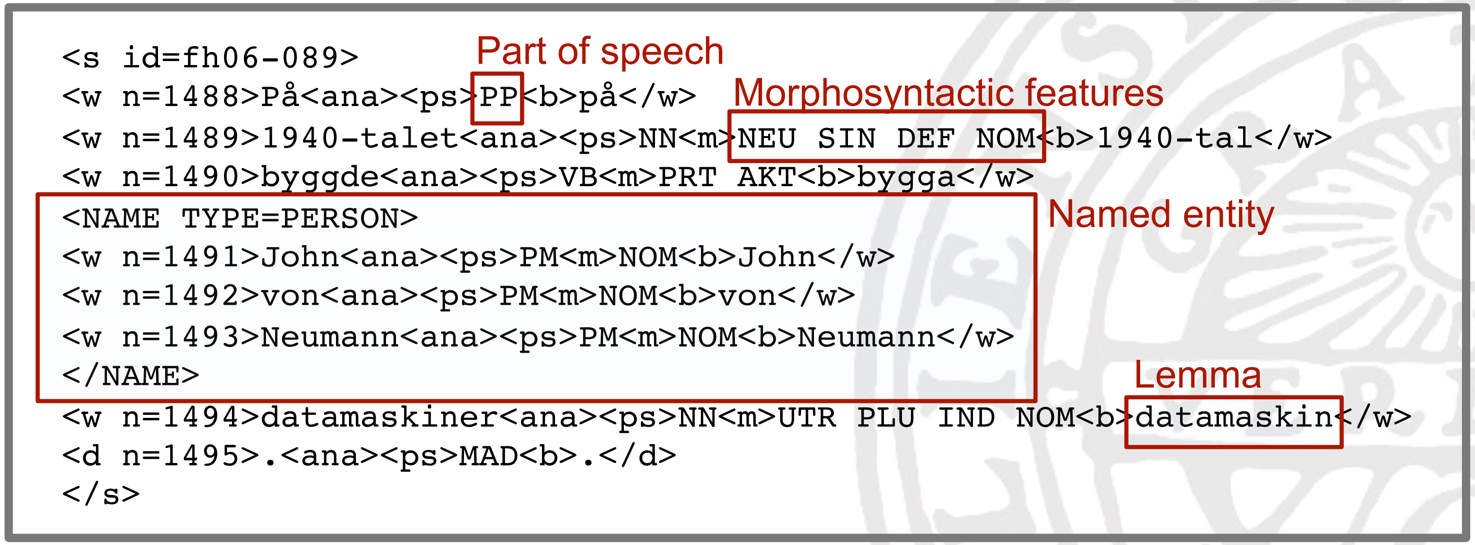
Lexical annotation

Syntactic annotation

```
*GENOM                      PR           AAPR
SKATTEREFORMEN              NNDDSS       AA
INFÖRS                      VVPSSMPA     FV
INDIVIDUELL                 AJ           SSAT
BESKATTNING                 VN           SS
AV                          PR           SSETPR
ARBETSINKOMSTER             NN   SS      SSET
.                           IP           IP
```

SUC

- Team led by Eva Ejerhed and Gunnel Källgren, 1990s
- Balanced corpus of written Swedish (1.2 million tokens)
- Annotation [Ejerhed et al. 1992]:
  - Parts of speech (PoS) + morphosyntactic features (MSF)
  - Lemmas
  - Named entities (SUC 2.0)



```
<s id=fh06-089>
<w n=1488>På<ana><ps>PP<b>på</w>
<w n=1489>1940-talet<ana><ps>NN<m>NEU SIN DEF NOM<b>1940-tal</w>
<w n=1490>byggde<ana><ps>VB<m>PRT AKT<b>bygga</w>
<NAME TYPE=PERSON>
<w n=1491>John<ana><ps>PM<m>NOM<b>John</w>
<w n=1492>von<ana><ps>PM<m>NOM<b>von</w>
<w n=1493>Neumann<ana><ps>PM<m>NOM<b>Neumann</w>
</NAME>
<w n=1494>datamaskiner<ana><ps>NN<m>UTR PLU IND NOM<b>datamaskin</w>
<d n=1495>.<ana><ps>MAD<b>.</d>
</s>
```

Part of speech
Morphosyntactic features
Named entity
Lemma

Methodology

Overall strategy:

- Keep SUC intact, modify Talbanken!
  - SUC is the larger corpus (minimize effort)
  - The SUC annotation scheme is a de facto standard
- Exception: Syntactic annotation

Major steps:

- Tokenization and sentence segmentation:
  - Make Talbanken conform to the principles of SUC
- Morphological annotation (PoS + MSF):
  - Reannotate Talbanken using a tagger trained on SUC
- Syntactic annotation:
  - Add phrase structure (PS) to Talbanken annotation
  - Annotate SUC using a parser trained on Talbanken
  - Derive dependency structure (DS) from PS+GF

UPPSALA
UNIVERSITET

Morphological Annotation

Reannotation of Talbanken:

- TnT tagger [Brants 2000]
- Self-training using SUC [Forsbom 2006]
- Estimated accuracy: 97.0%

Transverse manual validation:

- Function words by word form
- Content words by PoS category

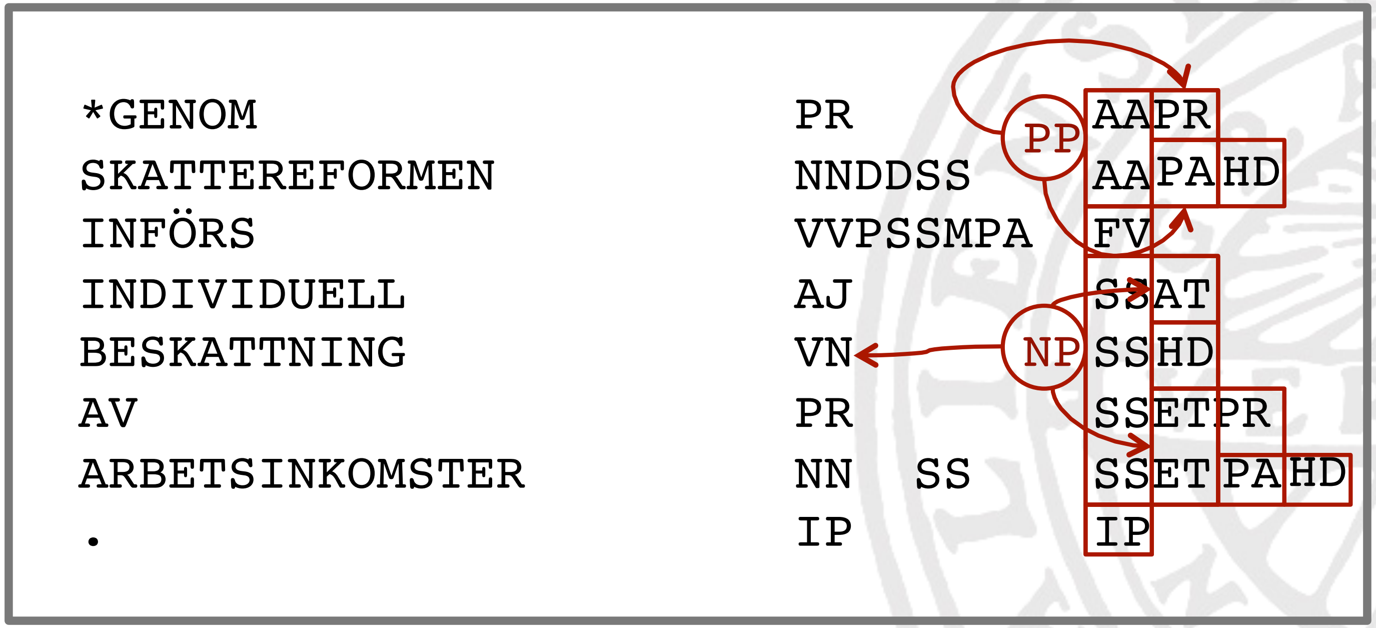
Speed-ups thanks to old annotation:

- Ambiguous forms: *men* (366 KN, 1 NN)
- Inflection vs. derivation: AB/JJ

Syntactic Annotation

Step 1: Enriching the MAMBA annotation

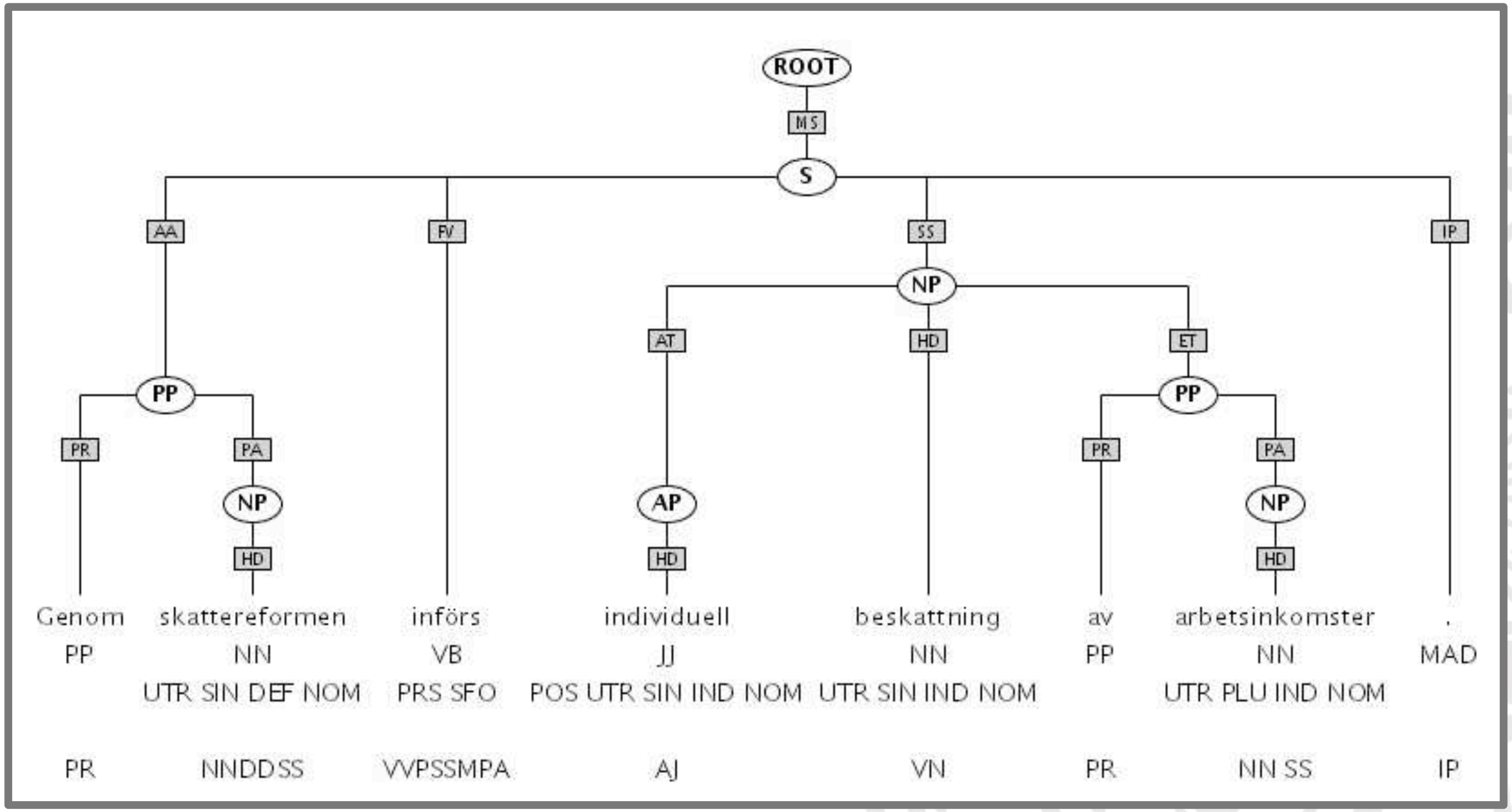
- Extract implicit PS+GF
- Insert additional structure (PP, VP, Coord)
- Infer nonterminal labels in PS



```
*GENOM              PR              AAPR
SKATTEREFORMEN      NNDDSS          AAPAHD
INFÖRS              VVPSSMPA        FV
INDIVIDUELL         AJ              SSAT
BESKATTNING         VN              SSHD
AV                  PR              SSETPR
ARBETSINKOMSTER     NN   SS         SSETPAHD
.                   IP              IP
```

Syntactic Annotation

The resulting PS+GF tree (Tiger-XML):



Syntactic Annotation

PS labels (8):

- ROOT, S, NP, VP, AP, AVP, PP, XP

GF labels (65):

- Predicate (4): end in V (verbal) or P (nonverbal)
- Subject (4): end in S; default SS
- Object (5): end in O; default OO
- Adverbial (12): end in A; default AA
- Coordination (4)
- Other GF (22)
- Punctuation (14)

Syntactic Annotation

Step 2: Parsing SUC

- MaltParser for PS+GF [Hall 2008a, 2008b]
- Trained on Talbanken's enriched annotation
- Estimated accuracy: 65% labeled F_1

Step 3: Validation

- Talbanken:
  - Manual correction of special test set (20k tokens)
- SUC:
  - Manual correction of special test set (20k tokens)
  - Automatic flagging of "suspicious structures"

UPPSALA
UNIVERSITET

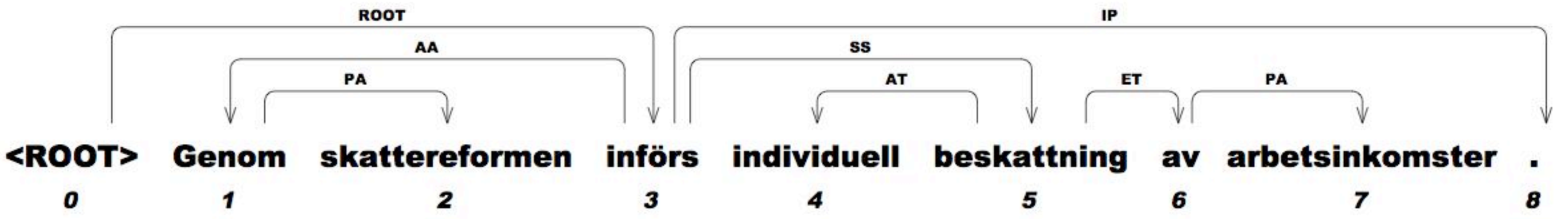
Syntactic Annotation

Step 4: Deriving dependency structures

- Structural conversion:
  - Head-finding rules based on GF labels:
    - If coordination, take conjunction (++) as head
    - Else use phrase-specific rules:
      - NP/AP/AVP: HD
      - S/VP: FV/VG/IV
      - PP: PR
  - Iterative refinement but no complete validation
- Labeling:
  - GF labels used as dependency labels

Syntactic Annotation

The resulting DS tree (CoNLL format):



```
1  Genom            _  PP   PP   _                      3  AA
2  skattereformen   _  NN   NN   UTR|SIN|DEF|NOM        1  PA
3  införs           _  VB   VB   PRS|SFO                0  ROOT
4  individuell      _  JJ   JJ   POS|UTR|SIN|IND|NOM    5  AT
5  beskattning      _  NN   NN   UTR|SIN|IND|NOM        3  SS
6  av               _  PP   PP   _                      5  ET
7  arbetsinkomster  _  NN   NN   UTR|PLU|IND|NOM        6  PA
8  .                _  MAD  MAD  _                      3  IP
```

Swedish Treebank 1.1

Layer	T [0.1M]	SUC [1.2M]
PoS+MSF	🟨	🟨
Lemma	⬛	🟨
PS+GF	⬜ 🟨	🟧 🟨
DS	⬜ ⬜	🟧 ⬜

Gold = manual validation

Silver = manual validation + conversion

Bronze = automatic annotation only

Parsing

Treebank Parsing

Goals:

- Develop better parsers (for Swedish)
- Compare different parsing architectures:
  - Representations (PS+GF vs. DS)
  - Modularization (tagging, parsing, labeling, …)
  - Models and algorithms

Fundamental view of parsing:

- Identify syntactic units and their relations
  - Phrases and grammatical functions in PS+GF
  - Heads and dependency relations in DS
  - Cross-framework evaluation?

UPPSALA
UNIVERSITET

Work in Progress

Dependency parsing (DS):

- Transition-based parsing (MaltParser)
- Impact of linguistic features
- Impact of training data (silver or bronze)

Phrase structure parsing (PS+GF):

- Treebank PCFGs
- Integration of function labels

Dependency Parsing

Transition-based parsing [Nivre 2008]:

- Transition system for deriving dependency trees
- Treebank-induced classifier for predicting transitions
- Parsing as greedy deterministic search

Basic setup:

- MaltParser 1.4.1 [http://maltparser.org]
- Transition system with online reordering [Nivre 2009]:
  - Ordinary shift-reduce parsing for projective trees
  - Permutation of word order for non-projective trees
  - Non-projective parsing in linear expected time
- Linear multi-class SVMs [Crammer and Singer 2001] using LIBLINEAR [Fan et al. 2008] for prediction

Feature Representation

[…, w_{-2}, w_{-1}, w_0]   [w_1, w_2, w_3, …]

[illegible]):

- [illegible] positions:
  - Dependency labels

    [illegible]

- [illegible] PoS:
  - [illegible]
- Trigrams:
  - Parts of speech in coordination

    (w_{-2}, w_{-1}, w_0), (w_{-1}, w_0, w_1), (w_0, w_1, w_2), (w_1, w_2, w_3)

- Leftmost and rightmost conjoined with PoS:

  w_{-1}, w_0

UPPSALA
UNIVERSITET

Feature Representation

Features	LAS	UAS
PoS	65.8	80.0
Dep	67.6	81.9
Lex	78.9	86.0
MSF	79.5	86.1
Dist	79.5	86.2
Prop	79.9	86.2

- Talbanken training set (5k sentences)
- 5-fold cross-validation
- Gold standard annotation as input (PoS, MSF)
- Labeled (LAS) and unlabeled (UAS) attachment score

Adding More Trees

Training Data	Talbanken	SUC
Talbanken (5k)	79.6	76.9
SUC-5k	74.8	73.3
SUC-75k	78.4	75.3
Talbanken + SUC-5k	79.1	76.3
Talbanken + SUC-75k	78.6	75.5

- Talbanken and SUC training sets
- Talbanken and SUC (development) test sets
- Gold standard annotation as input (PoS, MSF)
- Labeled (LAS) attachment score

Harvesting the Good Trees

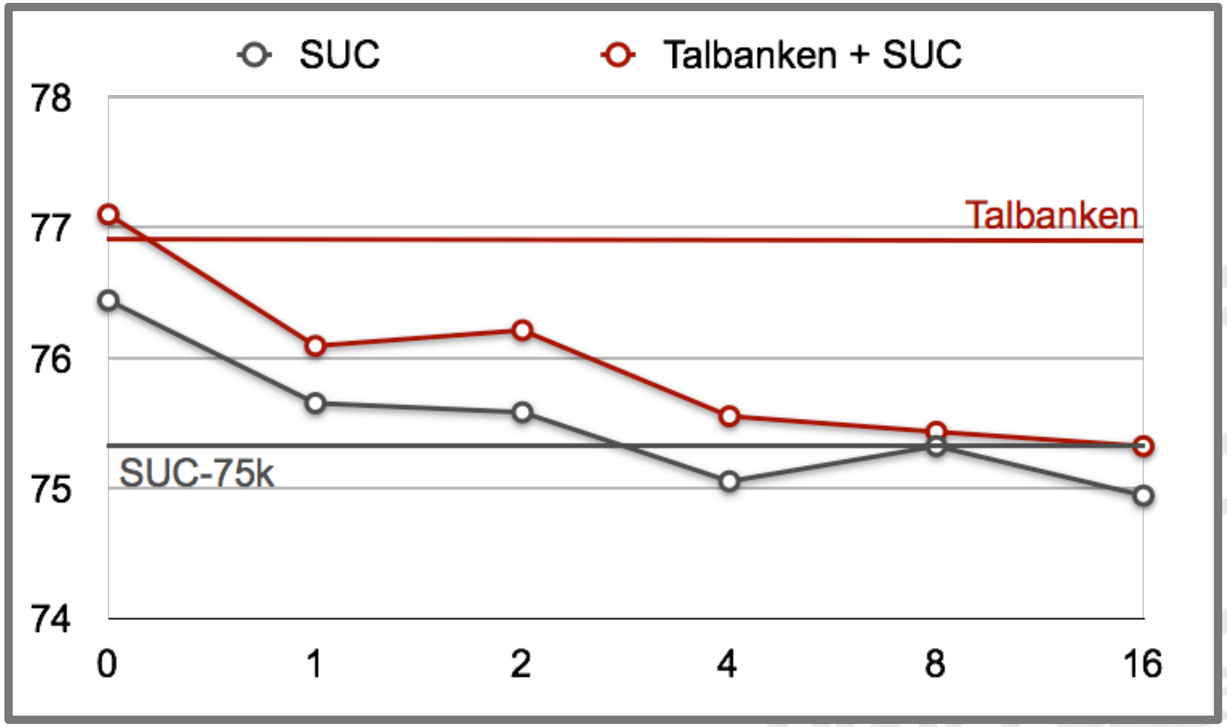
Warning flags:

- Automatic annotation of disallowed structures
- Substitute for manual revision in SUC

Eight flag categories:

Unary	Unary branching node
Nonterminal	Invalid PS label
Function	Invalid GF label
ForbiddenFunction	GF incompatible with PS/PoS
ForbiddenChild	Child with incompatible GF
ForbiddenSibling	Sibling with incompatible GFs
ObligatoryChild	Obligatory child GF missing
ObligatorySibling	Obligatory sibling GF missing

Harvesting the Good Trees



- SUC-42k training sets (with and without Talbanken)
- Random samples with at most *k* warning flags
- SUC (development) test set
- Labeled (LAS) attachment score

UPPSALA UNIVERSITET

Phrase Structure Parsing

Representation	Gold	Raw
PS	72.3	65.9
PS + GF	74.0	67.4
PS + parent annotation	74.6	68.4

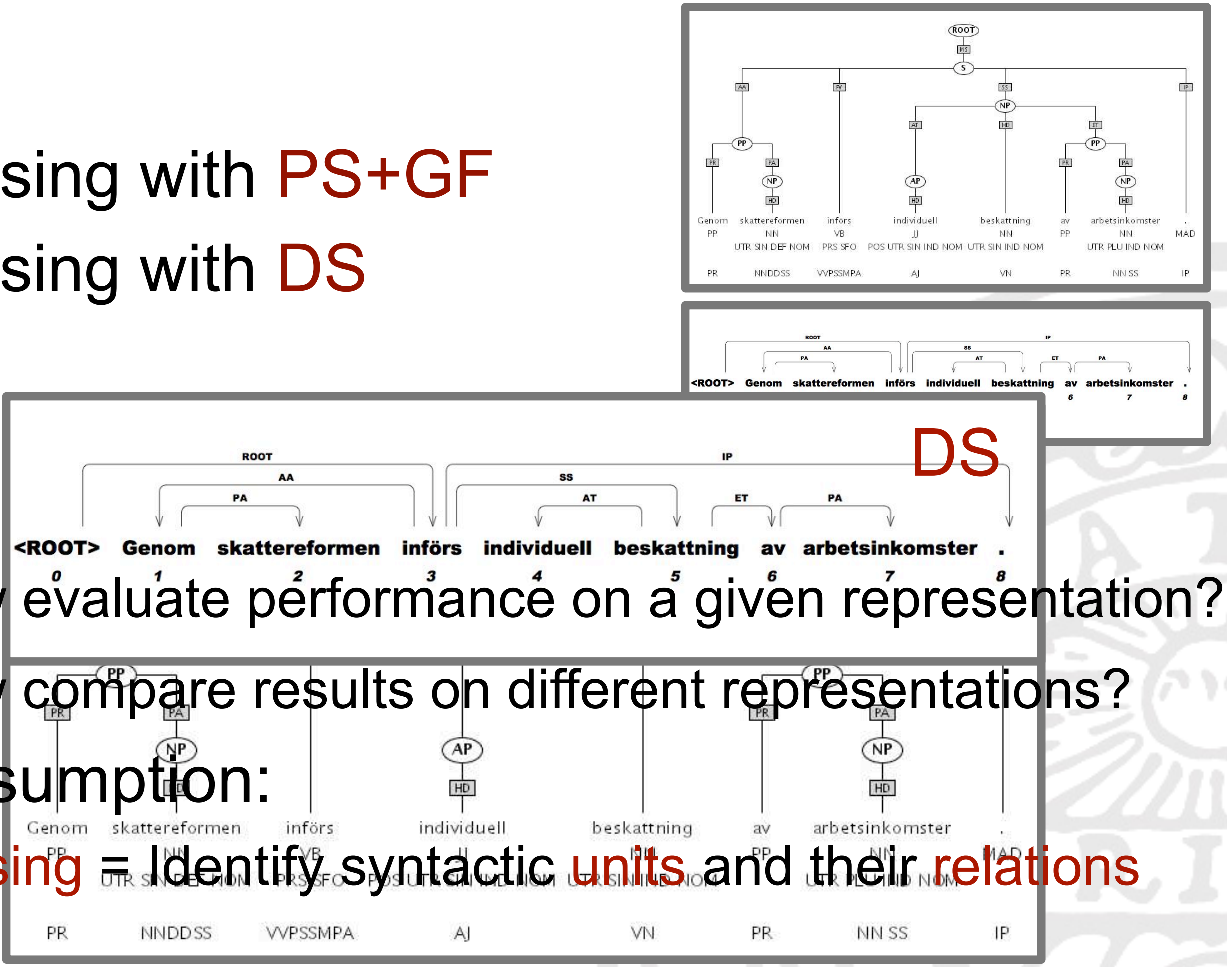
- Talbanken training set (5191 sentences)
- Talbanken (development) test sets
- Treebank PCFG (minimal smoothing)
- With and without gold standard annotation as input (PoS)
- PARSEVAL labeled F_1

Evaluation

UPPSALA
UNIVERSITET

The Problem

- Parsing with PS+GF
- Parsing with DS


ROOT
MS
S
AA
FV
SS
IP
PP
NP
AT
HD
ET
PR
PA
AP
Genom PP
skattereformen NN UTR SIN DEF NOM
införs VB PRS SFO
individuell JJ POS UTR SIN IND NOM
beskattning NN UTR SIN IND NOM
av PP
arbetsinkomster NN UTR PLU IND NOM
. MAD
PR NNDDSS VVPSSMPA AJ VN PR NN SS IP
DS
ROOT
IP
AA
SS
PA
AT
ET
PA
<ROOT> Genom skattereformen införs individuell beskattning av arbetsinkomster .
0 1 2 3 4 5 6 7 8

Issues:

- How evaluate performance on a given representation?
- How compare results on different representations?

Basic assumption:

- Parsing = Identify syntactic units and their relations

Cross-Framework Evaluation

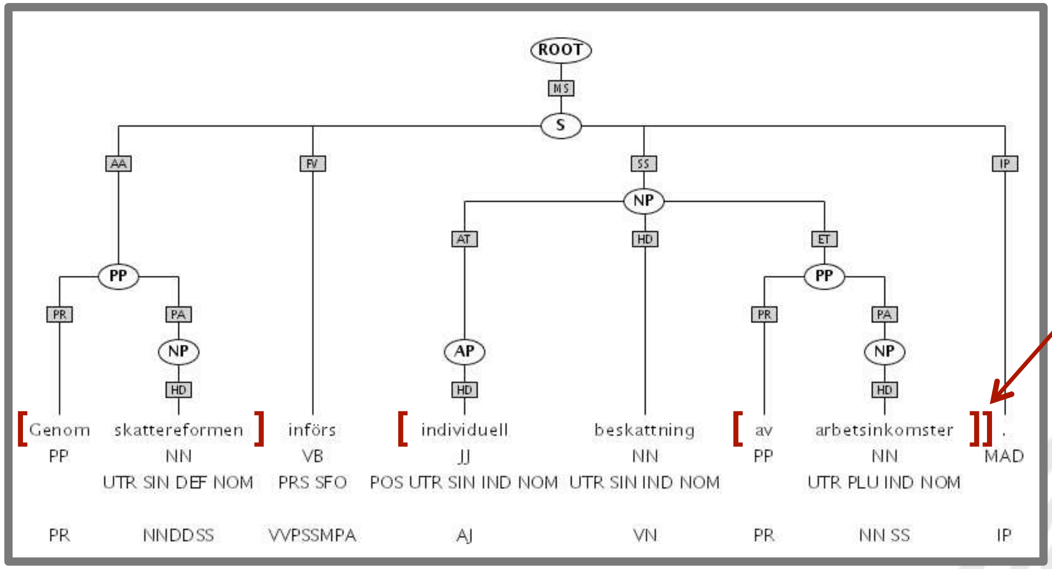
Two strategies:

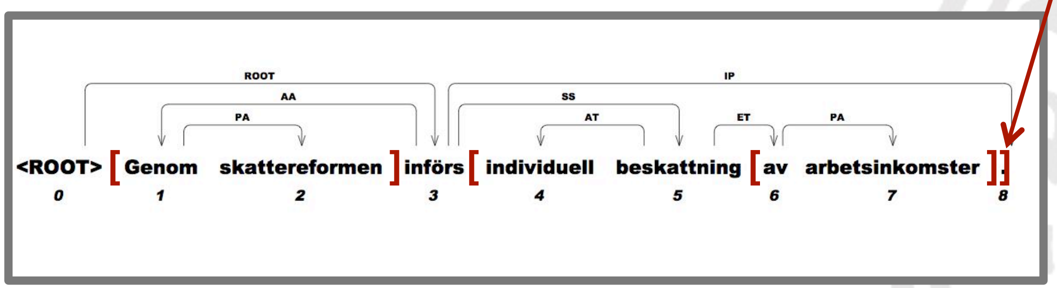
- Abstract over differences in representations
  - PARSEVAL [Black et al. 1991]
  - Problem: Metric may be uninformative (or misleading)
- Convert to (other) target representation
  - Labeled dependencies [Lin 1995, Carroll et al. 1998, Cer et al. 2010, Candito et al. 2010]
  - Problem: Conversions may be lossy

Our vision:

- Abstraction to target representation (almost)
- Informative without lossy conversion
- Evaluate capacity to recover units and relations

Spans




ROOT
AA
PA
IP
SS
AT
ET
PA
<ROOT> [Genom skattereformen] införs [individuell beskattning [av arbetsinkomster]] .
0 1 2 3 4 5 6 7 8

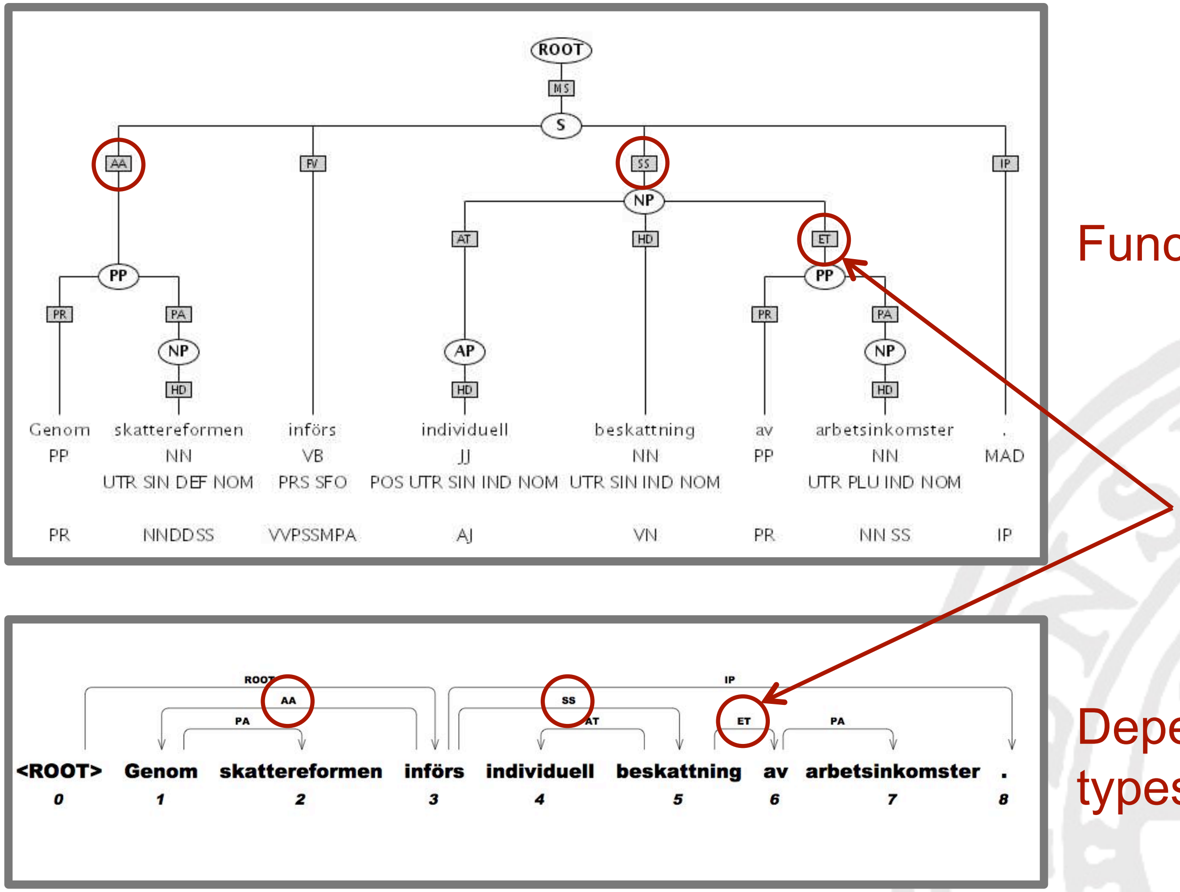
Spans

Brackets in PS

Subtree yields in DS

- No labels – abstraction over PS+GF

Relations



ROOT
MS
S
AA
FV
SS
IP
PP
NP
AT
HD
ET
PR
PA
AP
Genom PP PR
skattereformen NN UTR SIN DEF NOM NNDDSS
införs VB PRS SFO VVPSSMPA
individuell JJ POS UTR SIN IND NOM AJ
beskattning NN UTR SIN IND NOM VN
av PP PR
arbetsinkomster NN UTR PLU IND NOM NN SS
. MAD IP
ROOT
AA
PA
SS
AT
IP
ET
PA
<ROOT> Genom skattereformen införs individuell beskattning av arbetsinkomster .
0 1 2 3 4 5 6 7 8

Functions in GF

Relations

Dependency types in DS

- No heads – abstraction over DS

Putting It All Together

Relations of spans to larger spans:

[I think we have compared apples and oranges.]

Sbj[I], Prd[think], Obj[we have compared apples and oranges]

[we have compared apples and oranges]

Sbj[we], Prd[have compared], Obj[apples and oranges]

Abstraction over:

- Phrase types (not available in DS)
- Syntactic heads (not available in PS+GF)

Relation filtering allows further abstraction:

- Verb groups – main or auxiliary verb as head
- Coordination – no constraints on internal structure

Related Work

Like PARSEVAL:

- Evaluates bracketing of syntactic units
- Differences:
  - Adds relations between units
  - Allows functional filtering of units

Like dependency banks:

- Evaluates syntactic relations
- Differences:
  - Adds syntactic units (spans)
  - Minimizes the need for conversion

Conclusion

Harvest Time

Swedish Treebank 1.1:

- 1.3 million words of written Swedish
- Morphological annotation (gold)
- Syntactic annotation (gold, silver, bronze)

Next year's crop:

- Further enrichment of annotation
  - Lemmatization in Talbanken
  - Feature propagation to phrase level
- Parsing in multiple frameworks
- Cross-framework evaluation

UPPSALA
UNIVERSITET

References

Ahrenberg, L. (2007) LinES: An English-Swedish Parallel Treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*, 270–273.

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

Brants, T. (2000) TnT – a Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP)*.

Candito, M., Nivre, J., Denis, P. and Henestroza Anguiano, E. (2010) Benchmarking of Statistical Dependency Parsers for French. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING), Posters*, 108–116.

Carroll, J., Briscoe, E. and Sanfilippo, A. (1998) Parser Evaluation: A Survey and a New Proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* 447–454.

Cer, D., de Marneffe, M.-C., Jurafsky, D. and Manning, C. D. (2010) Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC).*

Crammer, K. and Singer, Y. (2001) On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research* 2, 265–292.

Einarsson, J. (1976a) Talbankens skriftspråkskonkordans. Lund University: Department of Scandinavian Languages.

References

Einarsson, J. (1976b) Talbankens talpråkskonkordans. Lund University: Department of Scandinavian Languages.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J. (2008) LIBLINEAR: A library for large linear classification *Journal of Machine Learning Research* 9, 1871–1874.

Forsbom, E. (2006) Big is Beautiful: Bootstrapping a PoS tagger for Swedish. Poster presentation at GSLT retreat, Gullmarsstrand, January 27–29, 2006.

Gustafson-Capková, S., Samuelsson, Y. and Volk, M. et al. (2007). SMULTRON (version 1.0) – The Stockholm MULtilingual parallel TReebank. http://www.ling.su.se/dali/research/smultron/index.htm. An English-German-Swedish parallel treebank with sub-sentential alignments.

Järborg, J. (1986) Manual för syntaggning. University of Gothenburg: Department of Swedish.

Kokkinakis, D. (2006) Towards a Swedish Medical Treebank. In Hajic, J. and Nivre, J. (eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, 199–210.

Lin, D. (1995) A Dependency-Based Method for Evaluating Broad-Coverage Parsers. In *Proceedings of IJCAI*, 1420–1425.

McDonald, R. (2006) Discriminative Training and Spanning Tree Algorithms for Dependency Parsing. PhD Thesis, University of Pennsylvania.

References

Megyesi, B., Dahlqvist, B., Pettersson, E. and Nivre, J. (2008) Swedish-Turkish Parallel Treebank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

Megyesi, B., Dahlqvist, B., Csato, E. A. and Nivre, J. (2010) The English-Swedish-Turkish Parallel Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*.

Nivre, J. (2008) Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics* 34(4), 513-553.

Nivre, J. (2009) Non-Projective Dependency Parsing in Expected Linear Time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 351-359.

Nivre, J., Nilsson, J. and Hall, J. (2006) Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 1392–1395.

Rayner, M., Carter, D., Bouillon, P., Digalakis, V. and Wirén, M. (2000) *The Spoken Language Translator*. Cambridge University Press.

Santamarta, L., Lindberg, N. and Gambäck, B. (1995) Towards Building a Swedish Treebank. In *Proceedings of the 10th Nordic Conference of Computational Linguistics*, 37–40.

Swedish Treebanking

Pioneering work:

- Talbanken [Einarsson 1976a, 1976b]
- SynTag [Järborg 1986]

More recent work:

- S-CLE [Santamarta et al. 1995, Rayner et al. 2000]
- Talbanken05 [Nivre et al. 2006]
- MEDLEX [Kokkinakis 2006]
- SMULTRON [Gustafson-Capková et al. 2007]
- LinES [Ahrenberg 2007]
- English-Swedish-Turkish Parallel Treebank [Megyesi et al. 2008, 2010]

Tokenization and Segmentation

Harmonization issues:

- Abbreviations and numerical expressions:
    - Always one token in SUC
    - Syntactically informed tokenization in Talbanken
- Sentence segmentation in lists:
    - Always one sentence per list item in SUC
    - Syntactically informed segmentation in Talbanken

Modifications implemented:

- Talbanken converted to SUC principles
- Completely automatic procedure

UPPSALA
UNIVERSITET

Morphological Annotation

Different tag sets in Talbanken and SUC:

	Talbanken	SUC
PoS tags	47	25
MSF tags	62	25
Complex tags	249	154

Incompatibilities:

- Different distinctions
- Different criteria of application
- No deterministic mapping possible

Part-of-Speech Categories

- Noun (NN)
- Proper noun (PM)
- Verb (VB)
- Participle (PC)
- Adjective (JJ)
- Adverb (AB)
- Wh-adverb (HA)
- Pronoun (PN)
- Wh-pronoun (HP)
- Possessive (PS)
- Wh-possessive (HS)
- Preposition (PP)
- Verb particle (PL)
- Determiner (DT)

- Wh-determiner (HD)
- Conjunction (KN)
- Subjunction (SN)
- Infinitive marker (IE)
- Cardinal numeral (RG)
- Ordinal numeral (RO)
- Interjection (PP)

- Major delimiter (MAD)
- Minor delimiter (MID)
- Paired delimiter (PAD)

- Foreign word (UO)

UPPSALA UNIVERSITET

Morphosyntactic Features

Verbs:

- Tense, Voice, Mood

Nouns and pronouns:

- Case, Definiteness, Gender, Number

Adjectives:

- Same as nouns + Comparison

Participles:

- Same as nouns + Tense

Adverbs:

- Comparison

All categories:

- Compound, Abbreviation

UPPSALA
UNIVERSITET

Swedish Treebank 1.1

Data Set	Sentences	Words	W/S
Talbanken training	4 941	75 970	15.4
Talbanken test	1 219	20 376	16.7
SUC training	72 674	1 143 274	15.7
SUC test	1 569	23 319	14.9

Statistics for different subsets of the Swedish Treebank:

- Number of sentences
- Number of words
- Average number of words per sentence

Changing the Parser

Training Data	Talbanken	SUC
Talbanken (5k)	79.6 (79.6)	74.9 (76.9)
SUC-5k	74.0 (74.8)	73.1 (73.3)
SUC-75k	77.7 (78.4)	75.1 (75.3)
Talbanken + SUC-5k	79.3 (79.1)	75.5 (76.3)
Talbanken + SUC-75k	79.5 (78.6)	75.4 (75.5)

- Talbanken and SUC training sets
- Talbanken and SUC (development) test sets
- Gold standard annotation as input (PoS, MSF)
- Labeled (LAS) attachment score
- MSTParser (2nd order, non-projective) [McDonald 2006]

Open Issues

Metrics:

- How define metrics for partial matches?
- Three types of errors:
  - Span
  - Relation
  - Domain (larger span)

Spans:

- Flat vs. deeply nested structures
- Incompatible spans

Relations:

- Recovery of relations for syntactic heads
- Long-distance dependencies