How to Compare Treebanks

Sandra Kübler, Wolfgang Maier, Ines Rehbein & Yannick Versley

LREC, May 2008
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 resources (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 resources (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:
  - Küberl (2005), Maier (2006), Küberl 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
Motivation

Parsing Experiments

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
Motivation
Parsing Experiments

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

Kübler, Maier, Rehbein & Versley
How to Compare Treebanks
The Treebanks: TiGer and TüBa-D/Z

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

Differences in annotation

<table>
<thead>
<tr>
<th></th>
<th>TiGer</th>
<th>TüBa-D/Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation</td>
<td>flat</td>
<td>more hierarchical</td>
</tr>
<tr>
<td>LDD</td>
<td>crossing branches</td>
<td>grammatical functions</td>
</tr>
<tr>
<td>Unary nodes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Topological fields</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
But without the Tigers will it no peace give

“But without the Tigers there will be no peace.”
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)**

<table>
<thead>
<tr>
<th></th>
<th>TiGer</th>
<th></th>
<th>TüBa-D/Z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bit</td>
<td>Lop</td>
<td>Stan</td>
<td>Bit</td>
</tr>
<tr>
<td>evalb</td>
<td>74.0</td>
<td>75.2</td>
<td>77.3</td>
<td>83.4</td>
</tr>
<tr>
<td>LA</td>
<td>90.9</td>
<td>91.3</td>
<td>92.4</td>
<td>91.5</td>
</tr>
</tbody>
</table>

- `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`
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 \(\rightarrow\) 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)

<table>
<thead>
<tr>
<th>evalb</th>
<th>TiGer Bit</th>
<th>Lop</th>
<th>Stan</th>
<th>TüBa-D/Z Bit</th>
<th>Lop</th>
<th>Stan</th>
</tr>
</thead>
<tbody>
<tr>
<td>74.0</td>
<td>75.2</td>
<td>77.3</td>
<td></td>
<td>83.4</td>
<td>84.6</td>
<td>88.5</td>
</tr>
<tr>
<td>90.9</td>
<td>91.3</td>
<td>92.4</td>
<td></td>
<td>91.5</td>
<td>91.8</td>
<td>93.6</td>
</tr>
</tbody>
</table>

### Labeled/unlabeled dependency accuracy (2000 sentences)

<table>
<thead>
<tr>
<th></th>
<th>TiGer Bit</th>
<th>Lop</th>
<th>Stan</th>
<th>TüBa-D/Z Bit</th>
<th>Lop</th>
<th>Stan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labelled Accuracy</td>
<td>78.8</td>
<td>80.5</td>
<td>81.6</td>
<td>71.3</td>
<td>72.8</td>
<td>75.9</td>
</tr>
<tr>
<td>Unlabelled Accuracy</td>
<td>83.0</td>
<td>84.5</td>
<td>85.6</td>
<td>81.7</td>
<td>83.4</td>
<td>86.8</td>
</tr>
</tbody>
</table>
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

Testing Parser Performance on Complex Grammatical Constructions
- 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
  ⇒ Criteria needed to evaluate grammatical constructions across treebanks
How to ensure inter-annotator agreement and reliability of human evaluation?
⇒ Error classification: describe categories for possible parser errors

**Example: Extraped RelativeClauses**

<table>
<thead>
<tr>
<th>Error description</th>
<th>TiGer</th>
<th>TüBa-D/Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Clause not recognized as relative clause</td>
<td>Grammatical function incorrect</td>
<td></td>
</tr>
<tr>
<td>(B) Head noun incorrect</td>
<td>Attachment error</td>
<td>SIMPX label instead of R-SIMPX</td>
</tr>
<tr>
<td>(C) Clause not recognized</td>
<td>Clause not recognized</td>
<td>Grammatical function incorrect</td>
</tr>
<tr>
<td>(D) Clause boundaries not correct</td>
<td>Span error</td>
<td>Clause not recognized</td>
</tr>
</tbody>
</table>
Results for Human Evaluation

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

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?


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


Motivation

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


Experimental Setup

Constituent Evaluation

Dependency Evaluation

Human Evaluation

Kübler, Maier, Rehbein & Versley

How to Compare Treebanks

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.


“But without the Tigers there will be no peace.”
But without the Tigers will it no peace give

“But without the Tigers there will be no peace.”

Kübler, Maier, Rehbein & Versley

How to Compare Treebanks
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)

<table>
<thead>
<tr>
<th></th>
<th>TiGer Bit</th>
<th>TiGer Lop</th>
<th>TiGer Stan</th>
<th>TüBa-D/Z Bit</th>
<th>TüBa-D/Z Lop</th>
<th>TüBa-D/Z Stan</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJ</td>
<td>80.2</td>
<td>81.1</td>
<td>78.7</td>
<td>74.6</td>
<td>75.3</td>
<td>76.1</td>
</tr>
<tr>
<td>OBJA</td>
<td>55.6</td>
<td>58.4</td>
<td>59.5</td>
<td>42.4</td>
<td>45.8</td>
<td>52.9</td>
</tr>
<tr>
<td>OBJD</td>
<td>11.6</td>
<td>11.5</td>
<td>14.1</td>
<td>12.9</td>
<td>13.3</td>
<td>13.1</td>
</tr>
<tr>
<td>PP</td>
<td>71.1</td>
<td>72.2</td>
<td>78.2</td>
<td>68.1</td>
<td>69.1</td>
<td>75.6</td>
</tr>
<tr>
<td>clause-sub.</td>
<td>57.0</td>
<td>58.2</td>
<td>60.9</td>
<td>45.8</td>
<td>47.5</td>
<td>52.1</td>
</tr>
</tbody>
</table>
Das Ziel sei es, “eine legale Organisation zu schaffen”, die unter anderem auch für die Medien ansprechbar sein soll. The goal be it, “a legal organisation to create”, which amongst others also ought to be approachable for the media.

“The aim is to create a legal organisation which, amongst others, also ought to be approachable for the media.”
(2) 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

<table>
<thead>
<tr>
<th></th>
<th>TiGer</th>
<th></th>
<th>TüBa-D/Z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bit</td>
<td>Lop</td>
<td>Stan</td>
<td>Bit</td>
</tr>
<tr>
<td>LAS ERC</td>
<td>76.2</td>
<td>76.0</td>
<td>77.4</td>
<td>71.6</td>
</tr>
<tr>
<td>FCR</td>
<td>79.5</td>
<td>74.4</td>
<td>81.8</td>
<td>78.5</td>
</tr>
<tr>
<td>PPN</td>
<td>76.8</td>
<td>79.7</td>
<td>87.0</td>
<td>75.5</td>
</tr>
<tr>
<td>PPV</td>
<td>73.6</td>
<td>80.9</td>
<td>79.2</td>
<td>65.8</td>
</tr>
<tr>
<td>CUC</td>
<td>65.2</td>
<td>67.0</td>
<td>70.7</td>
<td>57.5</td>
</tr>
<tr>
<td>SGF</td>
<td>76.1</td>
<td>77.2</td>
<td>79.3</td>
<td>74.0</td>
</tr>
<tr>
<td>ALL</td>
<td>73.3</td>
<td>73.9</td>
<td>76.8</td>
<td>69.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TiGer</th>
<th></th>
<th>TüBa-D/Z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UAS ERC</td>
<td>Bit</td>
<td>Lop</td>
<td>Stan</td>
</tr>
<tr>
<td></td>
<td>81.1</td>
<td>80.8</td>
<td>82.0</td>
<td>79.1</td>
</tr>
<tr>
<td>FCR</td>
<td>82.7</td>
<td>77.8</td>
<td>85.6</td>
<td>85.4</td>
</tr>
<tr>
<td>PPN</td>
<td>84.2</td>
<td>86.4</td>
<td>89.3</td>
<td>84.8</td>
</tr>
<tr>
<td>PPV</td>
<td>78.1</td>
<td>86.0</td>
<td>86.0</td>
<td>81.3</td>
</tr>
<tr>
<td>CUC</td>
<td>69.7</td>
<td>71.5</td>
<td>74.7</td>
<td>66.1</td>
</tr>
<tr>
<td>SGF</td>
<td>81.7</td>
<td>82.5</td>
<td>83.6</td>
<td>82.8</td>
</tr>
<tr>
<td>ALL</td>
<td>78.1</td>
<td>78.7</td>
<td>81.0</td>
<td>78.3</td>
</tr>
</tbody>
</table>

Labeled/unlabeled dependency accuracy for the testsuite