Old School vs. New School: Comparing Transition-Based Parsers with and without Neural Network Enhancement

Miryam de Lhoneux, Sara Stymne & Joakim Nivre

UPPSALA UNIVERSITY

TLT15, 20 January 2016
Overview

1. Introduction

2. Treebank Sampling

3. Comparing Parsing Accuracy

4. Impact of Training Size on Neural Network Parsing

5. Error Analysis

6. Conclusion
Introduction

Motivation

UD treebanks are growing in number and size at a fast pace, making our models generalisable across languages and domains. Parsing models enhanced by neural networks have seen a large boost in accuracy but are expensive to optimise.

Our proposal: work on a small but representative sample of UD treebanks.

Miryam de Lhoneux, Sara Stymne & Joakim Nivre
Motivation

- UD treebanks
  - are growing in number and size at a fast pace
  - make our models generalisable across languages and domains
Introduction

Motivation

- UD treebanks
  - are growing in number and size at a fast pace
  - make our models generalisable across languages and domains
- Parsing models enhanced by neural networks
  - have seen a large boost in accuracy
  - are expensive to optimise
Introduction

Motivation

- UD treebanks
  - are growing in number and size at a fast pace
  - make our models generalisable across languages and domains
- Parsing models enhanced by neural networks
  - have seen a large boost in accuracy
  - are expensive to optimise

Our proposal

Work on a small but representative sample of UD treebanks
Introduction

Motivation

Straka et al. (2015) trained Parsito, a neural network parser for UD. Limited comparison with MaltParser (Nivre et al., 2007).

This Work

Carry out a comparison of the two on a small sample.
Introduction

Motivation

Straka et al. (2015) trained Parsito, a neural network parser for UD Limited comparison with MaltParser (Nivre et al., 2007).
Introduction

Motivation
Straka et al. (2015) trained Parsito, a neural network parser for UD Limited comparison with MaltParser (Nivre et al., 2007).

This Work
Carry out a comparison of the 2 on a small sample
Treebank Sampling

Selection criteria

- Typical variety: 8 different fine-grained and 5 coarse-grained families
- Isolating, morphologically-rich and inflecting languages
- Variety in treebanks sizes and domains
- Availability of morphological features
- Quality of the treebank (according to UD validation tests)
**Selection criteria**

- **typological variety:**

  - 8 different fine-grained and 5 coarse-grained families
  - Isolating, morphologically-rich and inflecting languages
  - Non-projectivity
  - Variety in treebank sizes and domains
  - Availability of morphological features
  - Quality of the treebank (according to UD validation tests)
Selection criteria

- typological variety:
  - 8 different fine-grained and 5 coarse-grained families
Treebank Sampling

Selection criteria

- typological variety:
  - 8 different fine-grained and 5 coarse-grained families
  - isolating, morphologically-rich and inflecting languages
Treebank Sampling

Selection criteria

- typological variety:
  - 8 different fine-grained and 5 coarse-grained families
  - isolating, morphologically-rich and inflecting languages
  - non-projectivity
Selection criteria

- **typological variety:**
  - 8 different fine-grained and 5 coarse-grained families
  - isolating, morphologically-rich and inflecting languages
  - non-projectivity
- **variety in treebanks sizes and domains**
Treebank Sampling

Selection criteria

- typological variety:
  - 8 different fine-grained and 5 coarse-grained families
  - isolating, morphologically-rich and inflecting languages
  - non-projectivity

- variety in treebanks sizes and domains

- availability of morphological features
Selection criteria

- **typological variety:**
  - 8 different fine-grained and 5 coarse-grained families
  - isolating, morphologically-rich and inflecting languages
  - non-projectivity

- **variety in treebanks sizes and domains**

- **availability of morphological features**

- **quality of the treebank (according to UD validation tests)**
### Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>AncientGreekPR</td>
<td>indo-eur</td>
<td>helenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

**Table 1**: Treebank Sample
## Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>Ancient Greek PR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

**Table 1**: Treebank Sample
## Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td><strong>Chinese</strong></td>
<td><strong>sino-tibetan</strong></td>
<td><strong>sinitic</strong></td>
<td><strong>isolating language</strong></td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>AncientGreekPR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

**Table 1**: Treebank Sample
### Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td><strong>Finnish</strong></td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>AncientGreekPR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

**Table 1:** Treebank Sample
<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td><strong>English</strong></td>
<td><strong>indo-eur</strong></td>
<td><strong>germanic</strong></td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>AncientGreekPR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

**Table 1**: Treebank Sample
## Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td><strong>Ancient Greek</strong></td>
<td><strong>PRindo-eur</strong></td>
<td><strong>hellenic</strong></td>
<td><strong>amount of non-projectivity</strong></td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

*Table 1*: Treebank Sample
### Table 1: Treebank Sample

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>Ancient Greek PR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td><strong>smallest; full manual check</strong></td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

Table 1: Treebank Sample
## Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>AncientGreekPR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiatic</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

*Table 1: Treebank Sample*
## Treebank Sampling

<table>
<thead>
<tr>
<th>language</th>
<th>coarse</th>
<th>fine</th>
<th>main argument for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>indo-eur</td>
<td>balto-slavic</td>
<td>largest UD treebank</td>
</tr>
<tr>
<td>Chinese</td>
<td>sino-tibetan</td>
<td>sinitic</td>
<td>isolating language</td>
</tr>
<tr>
<td>Finnish</td>
<td>uralo-altaic</td>
<td>finno-ungric</td>
<td>many different domains</td>
</tr>
<tr>
<td>English</td>
<td>indo-eur</td>
<td>germanic</td>
<td>largest + full manual check</td>
</tr>
<tr>
<td>Ancient Greek PR</td>
<td>indo-eur</td>
<td>hellenic</td>
<td>amount of non-projectivity</td>
</tr>
<tr>
<td>Kazakh</td>
<td>uralo-altaic</td>
<td>turkic</td>
<td>smallest; full manual check</td>
</tr>
<tr>
<td>Tamil</td>
<td>dravidian</td>
<td>tamil</td>
<td>small; language family</td>
</tr>
<tr>
<td>Hebrew</td>
<td>afro-asiant</td>
<td>semitic</td>
<td>language family</td>
</tr>
</tbody>
</table>

Table 1: Treebank Sample
Comparing Parsing Accuracy

UDPipe (Straka et al., 2016):
Pretrained models Optimised using UDPipe's random hyperparameter search

MaltParser:
Arc-standard swap system with lazy oracle Feature model optimised with MaltOptimizer (Ballesteros and Nivre, 2016).

UDPipe used for tagging POS and morphological features

SyntaxNet (Andor et al., 2016):
Comparing Parsing Accuracy

UDPipe (Straka et al., 2016):
- Pretrained models
- Optimised using UDPipe’s random hyperparameter search
Comparing Parsing Accuracy

UDPipe (Straka et al., 2016):
- Pretrained models
- Optimised using UDPipe’s random hyperparameter search

MaltParser:
- Arc-standard swap system with lazy oracle
- Feature model optimised with MaltOptimizer (Ballesteros and Nivre, 2016).
Comparing Parsing Accuracy

UDPipe (Straka et al., 2016):
- Pretrained models
- Optimised using UDPipe’s random hyperparameter search

MaltParser:
- Arc-standard swap system with lazy oracle
- Feature model optimised with MaltOptimizer (Ballesteros and Nivre, 2016).

UDPipe used for tagging POS and morphological features
Comparing Parsing Accuracy

UDPipe (Straka et al., 2016):
- Pretrained models
- Optimised using UDPipe’s random hyperparameter search

MaltParser:
- Arc-standard swap system with lazy oracle
- Feature model optimised with MaltOptimizer (Ballesteros and Nivre, 2016).

UDPipe used for tagging POS and morphological features

SyntaxNet (Andor et al., 2016):
- *reported results*
Comparing Parsing Accuracy
Comparing Parsing Accuracy

![Graph comparing parsing accuracy for various languages with MALT and UDPipe.](image)
Comparing Parsing Accuracy

The diagram compares parsing accuracy for various languages using three different tools: MALT, UDPipe, and SyntaxNet. The accuracy is measured in LAS (Linearly Interpolated Synchronous Accuracy). The languages included are Czech, English, Greek, Finnish, Chinese, Hebrew, Tamil, and Kazakh, with sample sizes ranging from 8K to 1503K.

- **Czech**: 1503K, MALT 80, UDPipe 82, SyntaxNet 82
- **English**: 254K, MALT 80, UDPipe 80, SyntaxNet 80
- **Greek**: 206K, MALT 80, UDPipe 80, SyntaxNet 80
- **Finnish**: 181K, MALT 80, UDPipe 80, SyntaxNet 80
- **Chinese**: 123K, MALT 80, UDPipe 80, SyntaxNet 80
- **Hebrew**: 115K, MALT 80, UDPipe 80, SyntaxNet 80
- **Tamil**: 8K, MALT 80, UDPipe 80, SyntaxNet 80
- **Kazakh**: 4K, MALT 80, UDPipe 80, SyntaxNet 80
Learning curve experiment:
Impact of Training Size on Neural Network Parsing

Learning curve experiment:

MaltParser: arc-standard swap with lazy oracle and extended feature model
UDPipe: arc-standard swap with lazy oracle and default hyperparameters
Learning curve experiment:

MaltParser: arc-standard swap with lazy oracle and extended feature model
UDPipe: arc-standard swap with lazy oracle and default hyperparameters

Splits: 1K to max with 50K splits
zoom on small data sizes: 1K to 15K
Impact of Training Size on Neural Network Parsing

Expectation

Accuracy

Training size

- NN parser
- non NN parser
Impact of Training Size on Neural Network Parsing

Reality

![Graphs showing the impact of training size on neural network parsing for different languages.](image-url)
Impact of Training Size on Neural Network Parsing

![Graph showing the impact of training size on neural network parsing. The x-axis represents the training size (1K, 51K, 101K, 151K) and the y-axis represents the parsing accuracy. The graph compares udpipe and maltparser. As the training size increases, the parsing accuracy also increases for both udpipe and maltparser.]
Impact of Training Size on Neural Network Parsing

![Graphs showing the impact of training size on neural network parsing for various languages including Czech, English, Chinese, Ancient Greek-PROIEL, Hebrew, Finnish, Tamil, and Kazakh.](image)
Inspired by McDonald and Nivre (2007): comparison of the accuracy of 2 parsers on a variety of graph and linguistics factors. Concatenating 9K of all development sets + all development set for Kazakh and Tamil.
Inspired by McDonald and Nivre (2007): comparison of the accuracy of 2 parsers on a variety of graph and linguistics factors
Inspired by McDonald and Nivre (2007): comparison of the accuracy of 2 parsers on a variety of graph and linguistics factors

Concatenating 9K of all development sets + all development set for Kazakh and Tamil
Error Analysis: Relation Length

![Graph showing F1 scores against relation length for udpipe and maltparser.](Image)
Error Analysis: Sentence Length

![Graph showing the LAS scores for two parsers, uaspipe and mältparser, across different sentence length bins. The graph shows a decline in LAS scores as sentence length increases.]
Error Analysis: Dependency Relations

Miryam de Lhoneux, Sara Stymne & Joakim Nivre

Old School vs. New School
Error Analysis: Dependency Relations

![Chart showing F1 scores for dependency relations with MALT and UDpipe]
Error Analysis: Dependency Relations

![Graph showing F1 scores for different dependency relations using Malt and UDPipe]

- Dependency Relations: nmod, punct, case, root, nsubj, dobj, det, amod, advmod, conj, cc, mark, nummod, det:df, advcl
- F1 scores are compared between Malt and UDPipe for each relation.
Error Analysis: Dependency Relations

![Graph showing F1 scores for different dependency relations using Malt and UDpipe]

- Dependency Relations: nmod, punct, case, root, nsubj, dobj, det, amod, advmod, conj, cc, mark, nummod, det:det, advcl

- Malt and UDpipe comparison for each dependency relation, with error bars indicating variability.
Error Analysis: POS tags

[Bar chart showing FL for different POS tags using 'malt' and 'udpipe' models.]
Error Analysis: POS tags

![Bar chart showing FL for different POS tags with error bars. The chart compares two systems: malt and udpipe. The chart includes POS tags such as NOUN, VERB, PUNCT, ADP, DET, PRON, ADJ, ADV, PROP, CONJ, PART, NUM, SCONJ, AUX, X, INTJ, and SYM. The y-axis represents FL ranging from 0 to 100, and the x-axis represents POS tags. The chart indicates that udpipe generally performs better than malt for most POS tags.]
Conclusion

UDPipe outperforms MaltParser with large training size.

The learning curve of UDPipe is steeper than MaltParser.

But flattens out in the same way as MaltParser.

MaltParser suffers more than UDPipe from longer sentences/dependencies.

Future Work

Effect of beam search

More fine-grained error analysis
Conclusion

UDPipe outperforms MaltParser with large training size.
The learning curve of UDPipe is steeper than MaltParser but flattens out in the same way as MaltParser.
MaltParser suffers more than UDPipe from longer sentences/dependencies.

Future Work

- Effect of beam search
- More fine-grained error analysis
Conclusion

- UDPipe outperforms MaltParser with large training size
Conclusion and Future Work

Conclusion

- UDPipe outperforms MaltParser with large training size
- The learning curve of UDPipe is steeper than MaltParser

Future Work

- Effect of beam search
- More fine-grained error analysis
Conclusion and Future Work

Conclusion

- UDPipe outperforms MaltParser with large training size
- The learning curve of UDPipe is steeper than MaltParser
- But flattens out in the same way as MaltParser

Future Work

- Effect of beam search
- More fine-grained error analysis
Conclusion

- UDPipe outperforms MaltParser with large training size
- The learning curve of UDPipe is steeper than MaltParser
- But flattens out in the same way as MaltParser
- MaltParser suffers more than UDPipe from longer sentences/dependencies
Conclusion

- UDPipe outperforms MaltParser with large training size
- The learning curve of UDPipe is steeper than MaltParser
- But flattens out in the same way as MaltParser
- MaltParser suffers more than UDPipe from longer sentences/dependencies

Future Work
Conclusion

- UDPipe outperforms MaltParser with large training size
- The learning curve of UDPipe is steeper than MaltParser
- But flattens out in the same way as MaltParser
- MaltParser suffers more than UDPipe from longer sentences/dependencies

Future Work

- Effect of beam search
Conclusion and Future Work

**Conclusion**
- UDPipe outperforms MaltParser with large training size
- The learning curve of UDPipe is steeper than MaltParser
- But flattens out in the same way as MaltParser
- MaltParser suffers more than UDPipe from longer sentences/dependencies

**Future Work**
- Effect of beam search
- More fine-grained error analysis


