# Non-projective Dependency-based Pre-Reordering with Recurrent Neural Network for Machine Translation

**Antonio Valerio Miceli Barone** Giuseppe Attardi

University of Pisa, Italy

Jun 4, 2015

# Pre-reordering for statistical machine translation

- Phrase-based machine translation quality is affected by the amount of reordering between languages [Birch et al. 2009]

# Pre-reordering for statistical machine translation

- Phrase-based machine translation quality is affected by the amount of reordering between languages [Birch et al. 2009]
- Pre-reordering source sentences improves quality

# Pre-reordering for statistical machine translation

- Phrase-based machine translation quality is affected by the amount of reordering between languages [Birch et al. 2009]
- Pre-reordering source sentences improves quality

die Katze hat die Frau gekauft .

*the woman has bought the cat .*

# Pre-reordering for statistical machine translation

- Phrase-based machine translation quality is affected by the amount of reordering between languages [Birch et al. 2009]
- Pre-reordering source sentences improves quality

die Katze hat die Frau gekauft .
**die Frau hat gekauft die Katze .**
*the woman has bought the cat .*

# Pre-reordering approaches

- Syntax-based hand-coded rules [Collins et al. 2005, ...]
  - ☺High quality improvement
  - ☹Require extensive language-specific linguistic expertise

# Pre-reordering approaches

- Syntax-based hand-coded rules [Collins et al. 2005, ...]
  - ☺High quality improvement
  - ☹Require extensive language-specific linguistic expertise
- Syntax-based statistical [Genzel, 2010, ...]
  - ☺Trained on word alignments, exploit syntax
  - ☹Only **tree-local** swaps, only constituency or **projective** dependency syntax

# Pre-reordering approaches

- Syntax-based hand-coded rules [Collins et al. 2005, ...]
  - ☺High quality improvement
  - ☹Require extensive language-specific linguistic expertise
- Syntax-based statistical [Genzel, 2010, ...]
  - ☺Trained on word alignments, exploit syntax
  - ☹Only **tree-local** swaps, only constituency or **projective** dependency syntax
- Syntax-free statistical [Tromble and Eisner, 2009, ...]
  - ☺Trained on word alignments
  - ☹Don't exploit syntax, only word pair features

# Example

auf diese Weise wird die raffinierte Undurchsichtigkeit geschickt aufrechterhalten ; während der Euro ″ stark und stabil ″ sein sollte und die Währungsreserven anfangs lediglich zur Verteidigung während des Übergangszeitraums ( falls notwendig ) dienen sollten , erweist sich heute , daßweder die eine noch die andere dieser Behauptungen zutreffend waren und sich in Frankfurt überhaupt nichts tut !

*the issue therefore remains skilfully blurred ; while the euro was intended to be ´ strong and stable ´ and the reserve assets were originally intended to provide protection during the transitional period ( should this prove necessary ) , it now appears that neither of these expectations has been fulfilled and Frankfurt is totally deadlocked !*

# Example

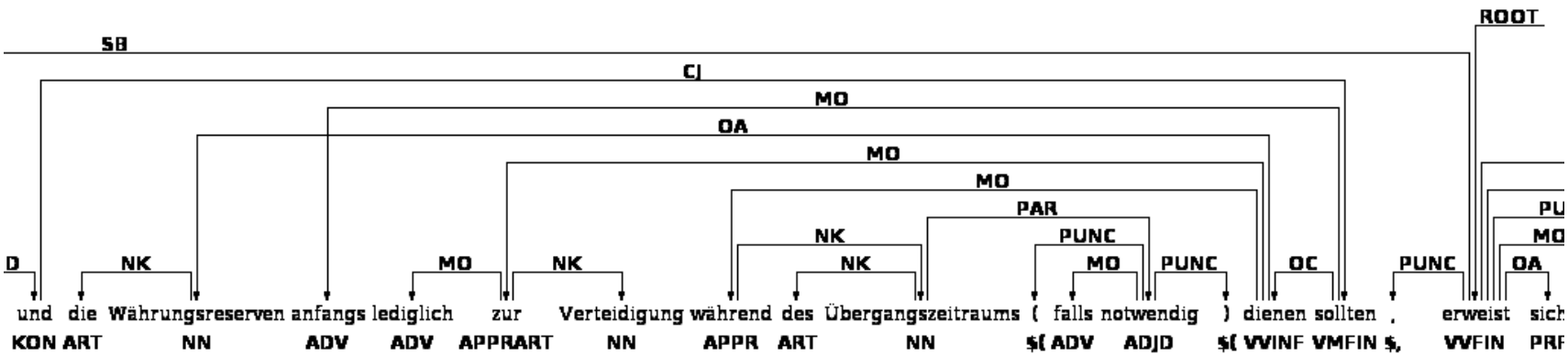
auf diese Weise wird die raffinierte Undurchsichtigkeit geschickt aufrechterhalten ; während der Euro ″ stark und stabil ″ sein sollte und die Währungsreserven anfangs lediglich zur Verteidigung während des Übergangszeitraums ( falls notwendig ) dienen sollten , erweist sich heute , daßweder die eine noch die andere dieser Behauptungen zutreffend waren und sich in Frankfurt überhaupt nichts tut !

**die raffinierte geschickt Undurchsichtigkeit aufrechterhalten ; während der Euro sollte auf sein ″ stark und stabil ″ und die Währungsreserven anfangs lediglich dienen sollten zur Verteidigung während des Übergangszeitraums ( diese Weise wird falls notwendig ) , erweist sich heute , daßweder die eine noch dieser Behauptungen zutreffend die andere waren und sich in Frankfurt überhaupt nichts tut !**

*the issue therefore remains skilfully blurred ; while the euro was intended to be ´ strong and stable ´ and the reserve assets were originally intended to provide protection during the transitional period ( should this prove necessary ) , it now appears that neither of these expectations has been fulfilled and Frankfurt is totally deadlocked !*

# Example

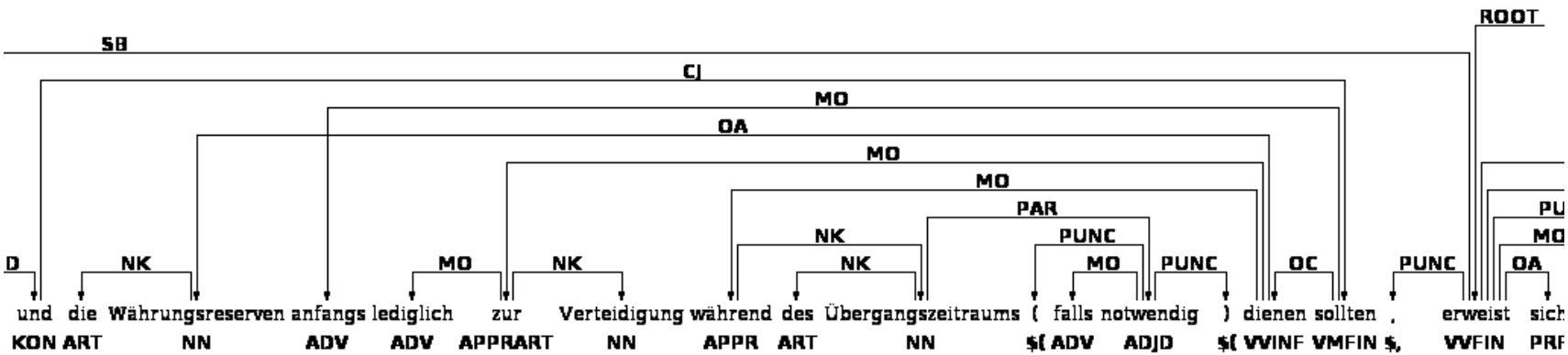
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




...

# Example

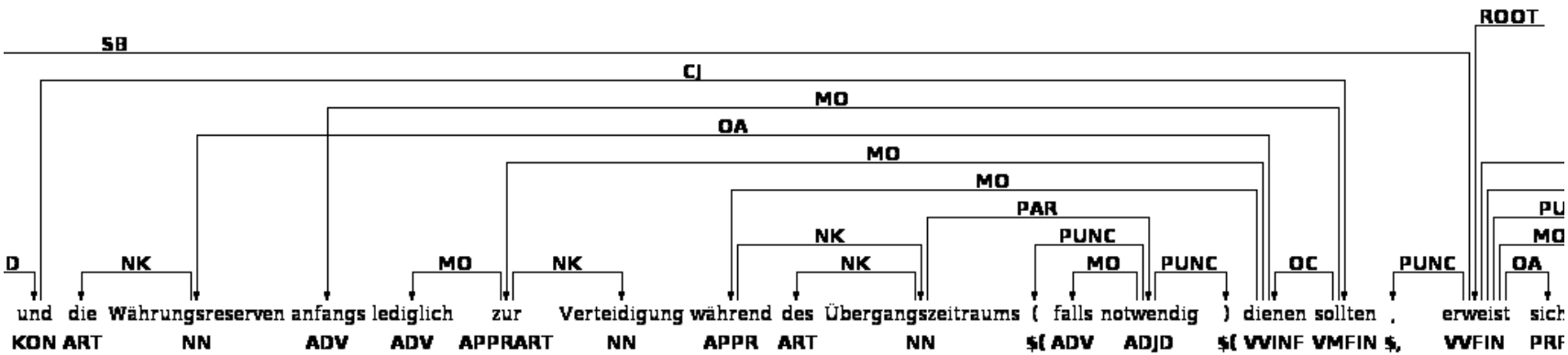
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die

# Example

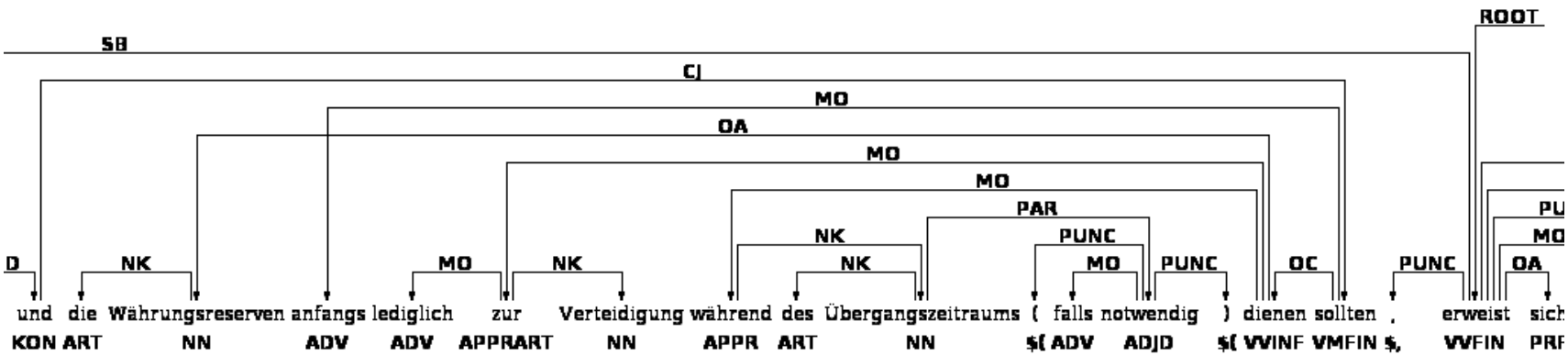
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven

# Example

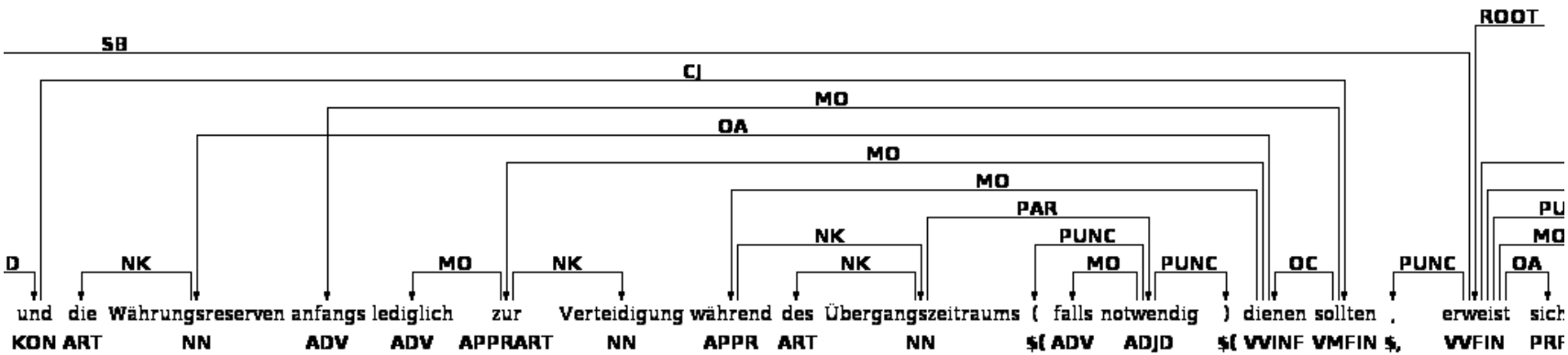
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven anfangs

# Example

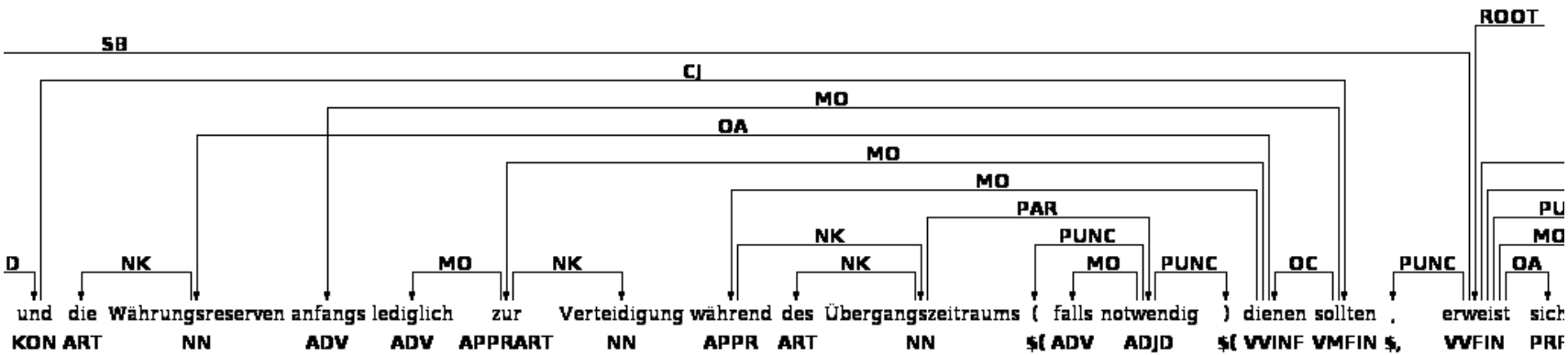
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven anfangs lediglich

# Example

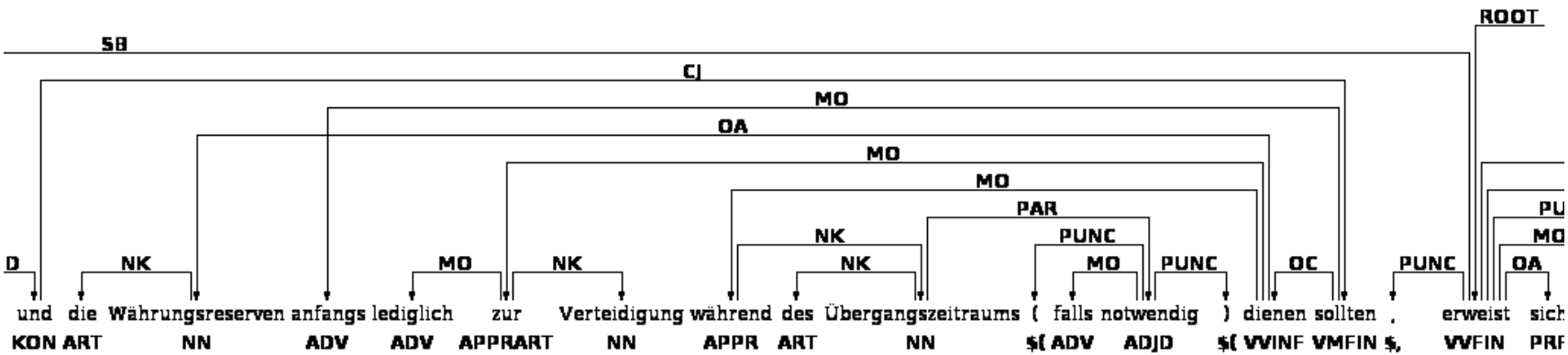
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven anfangs lediglich dienen

# Example

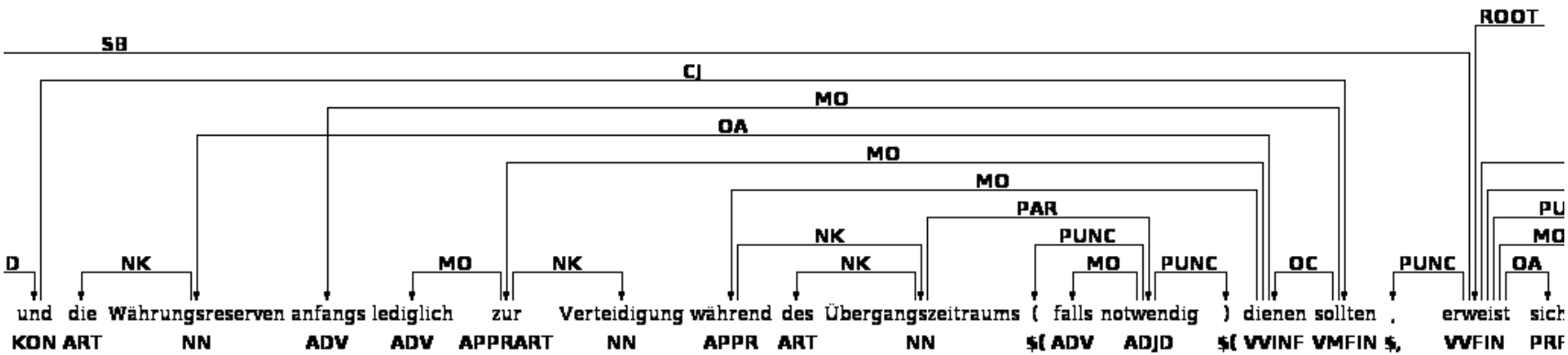
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven anfangs lediglich dienen sollten

# Example

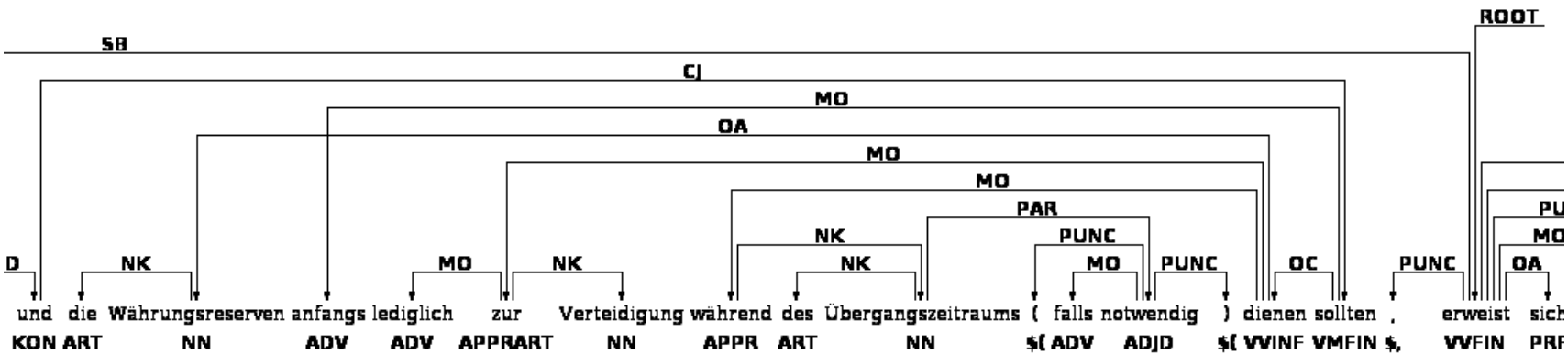
... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven anfangs lediglich dienen sollten zur

# Example

... die Währungsreserven anfangs lediglich zur Verteidigung ... dienen sollten ...
... *the reserve assets were originally intended to provide protection* ...




... die Währungsreserven anfangs lediglich dienen sollten zur Verteidigung ...

# Our proposal

- Arbitrary permutation prediction

# Our proposal

- Arbitrary permutation prediction
  - Use syntactical features from a **non-projective** dependency parse

# Our proposal

- Arbitrary permutation prediction
  - Use syntactical features from a **non-projective** dependency parse
  - Allow **non-tree-local** moves

# Our proposal

- Arbitrary permutation prediction
  - Use syntactical features from a **non-projective** dependency parse
  - Allow **non-tree-local** moves
- Leverage language models technology

# Our proposal

- Arbitrary permutation prediction
  - Use syntactical features from a **non-projective** dependency parse
  - Allow **non-tree-local** moves
- Leverage language models technology
  - ☺Tried and tested approach [Feng et al. 2010, ...]

# Our proposal

- Arbitrary permutation prediction
  - Use syntactical features from a **non-projective** dependency parse
  - Allow **non-tree-local** moves
- Leverage language models technology
  - ☺Tried and tested approach [Feng et al. 2010, ...]
  - ☹Positive probability mass to non-permutations

# Our proposal

- Arbitrary permutation prediction
  - Use syntactical features from a **non-projective** dependency parse
  - Allow **non-tree-local** moves
- Leverage language models technology
  - ☺Tried and tested approach [Feng et al. 2010, ...]
  - ☹Positive probability mass to non-permutations
  - ☹NNLMs have performance issues due to normalization over dictionary

# Recurrent Neural Network Reordering Model

- Take the Recurrent Neural Network Language Model [Mikolov et al. 2010]

# Recurrent Neural Network Reordering Model

- Take the Recurrent Neural Network Language Model [Mikolov et al. 2010]
- Change output soft-max to consider only unemitted words in the current sentence

# Recurrent Neural Network Reordering Model

- Take the Recurrent Neural Network Language Model [Mikolov et al. 2010]
- Change output soft-max to consider only unemitted words in the current sentence
- Include syntax features (POS, DEPREL, parent POS, ...)

# Recurrent Neural Network Reordering Model

- Take the Recurrent Neural Network Language Model [Mikolov et al. 2010]
- Change output soft-max to consider only unemitted words in the current sentence
- Include syntax features (POS, DEPREL, parent POS, ...)
- and permutation-specific word pair features (distortion distance, dependency relation ...)

# Recurrent Neural Network Reordering Model

- Take the Recurrent Neural Network Language Model [Mikolov et al. 2010]
- Change output soft-max to consider only unemitted words in the current sentence
- Include syntax features (POS, DEPREL, parent POS, ...)
- and permutation-specific word pair features (distortion distance, dependency relation ...)

$$
\begin{aligned}
v(t) &= \tau(\Theta^{(1)} \cdot x(t) + \Theta^{REC} \cdot v(t-1)) \\
h_j(t) &= < \tau(\Theta^{(o)} \cdot x_o(j)), \theta^{(2)} \odot v(t-1) > \\
&\quad + \theta^{(\alpha)} \cdot \log(L_f - t) + \theta^{(bias)} \qquad (1) \\
p_j(t) &= \frac{\exp h_j(t)}{\sum_{j'} \exp h_{j'}(t)}
\end{aligned}
$$

# Results

Trained on German-to-English Europarl v7

# Results

Trained on German-to-English Europarl v7

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |

# Results

Trained on German-to-English Europarl v7

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |

Translation BLEU:

| Test set | system | BLEU | improvement |
|---|---|---|---|
| Europarl | baseline | 33.00 | |
| Europarl | ”oracle” | 41.80 | +8.80 |
| Europarl | Collins | 33.52 | +0.52 |
| Europarl | unlex. Base RNN-RM | 33.41 | +0.41 |
| Europarl | lex. Base RNN-RM | 33.38 | +0.38 |
| news2013 | baseline | 18.80 | |
| news2013 | Collins | NA | NA |
| news2013 | unlex. Base RNN-RM | 19.19 | +0.39 |
| news2013 | lex. Base RNN-RM | 19.01 | +0.21 |
| news2009 | baseline | 18.09 | |
| news2009 | **Collins** | 18.74 | **+0.65** |
| news2009 | unlex. Base RNN-RM | 18.50 | +0.41 |
| news2009 | lex. Base RNN-RM | 18.44 | +0.35 |

# Reordering automaton

- Walk on dependency tree and emit word

# Reordering automaton

- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions

# Reordering automaton

- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions
  - Constraints to ensure proper permutation and no cycles

# Reordering automaton

- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions
  - Constraints to ensure proper permutation and no cycles
  - Each permutation bijectively maps to an automaton execution

# Reordering automaton

- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions
  - Constraints to ensure proper permutation and no cycles
  - Each permutation bijectively maps to an automaton execution
- Directly predict sequence of actions [Miceli-Barone and Attardi, 2013]

# Reordering automaton

- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions
  - Constraints to ensure proper permutation and no cycles
  - Each permutation bijectively maps to an automaton execution
- Directly predict sequence of actions [Miceli-Barone and Attardi, 2013]
  - ☹Didn't work well. Executions lengths are variable

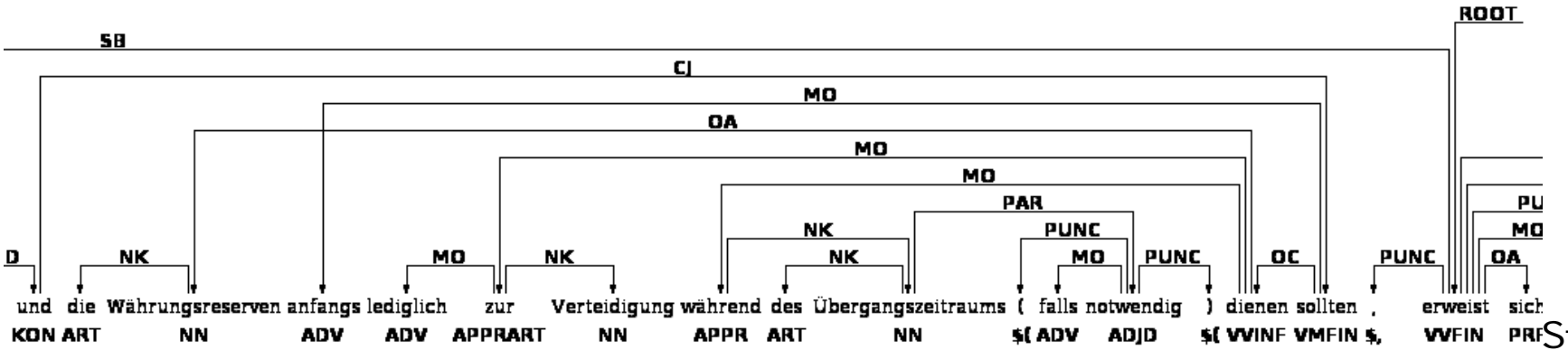
# Reordering automaton

- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions
  - Constraints to ensure proper permutation and no cycles
  - Each permutation bijectively maps to an automaton execution
- Directly predict sequence of actions [Miceli-Barone and Attardi, 2013]
  - ☹Didn't work well. Executions lengths are variable
- Or
  - Break sequence of actions into **fragments** at word emission boundaries

# Reordering automaton

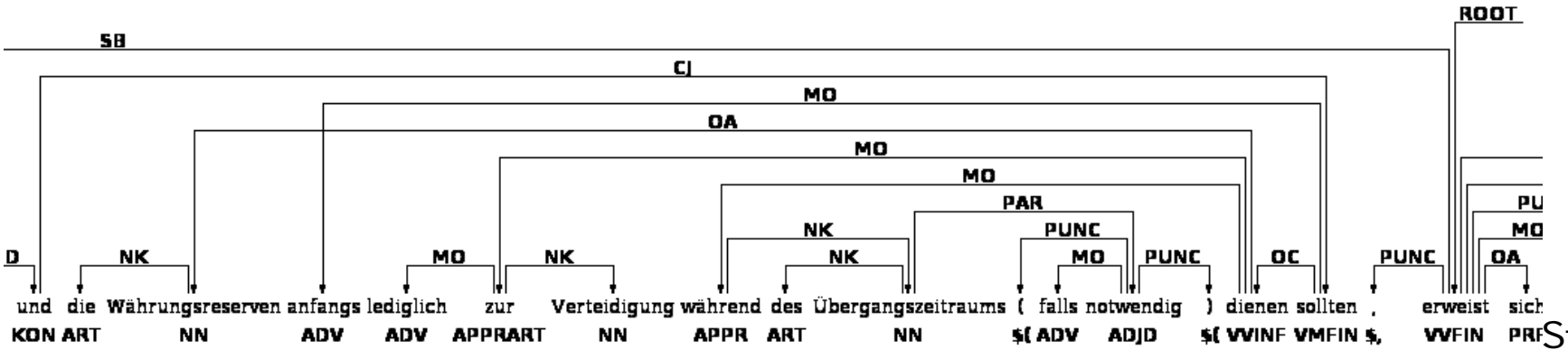
- Walk on dependency tree and emit word
  - *EMIT*, *UP* and $DOWN_{child}$ actions
  - Constraints to ensure proper permutation and no cycles
  - Each permutation bijectively maps to an automaton execution
- Directly predict sequence of actions [Miceli-Barone and Attardi, 2013]
  - ☹Didn't work well. Executions lengths are variable
- Or
  - Break sequence of actions into **fragments** at word emission boundaries
  - Predict permutation at word level, including fragments as features

# Reordering automaton example




Starting from **sollten**

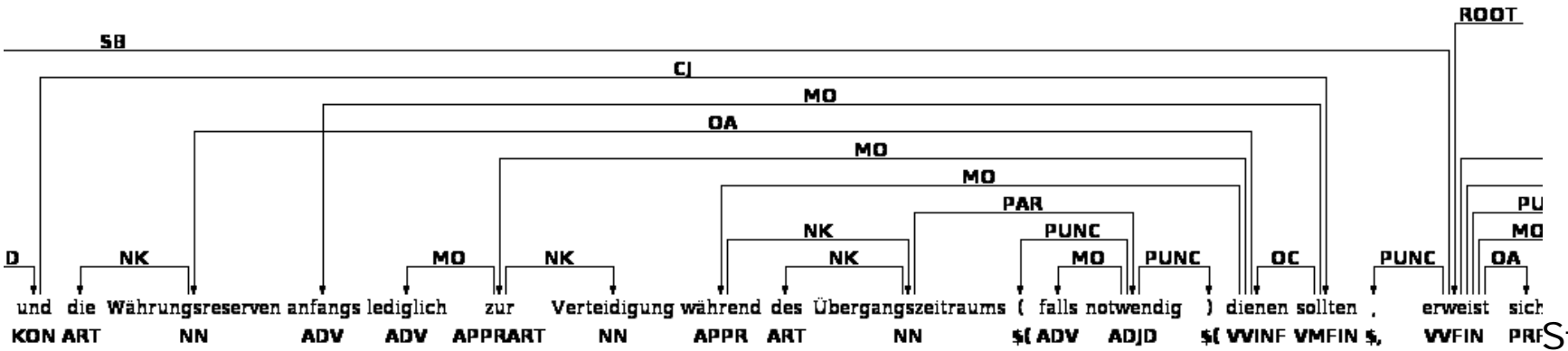
# Reordering automaton example




Starting from **sollten**

- $DOWN_{dienen}$ $DOWN_{Wahr...}$ $DOWN_{die}$ *EMIT* **die**

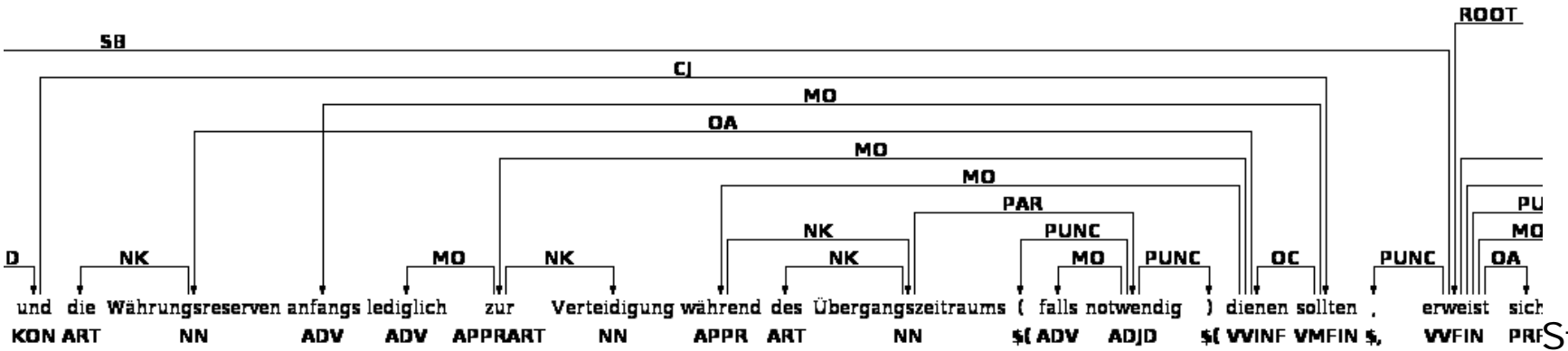
# Reordering automaton example




Starting from **sollten**

- $DOWN_{dienen}$ $DOWN_{Wahr...}$ $DOWN_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**

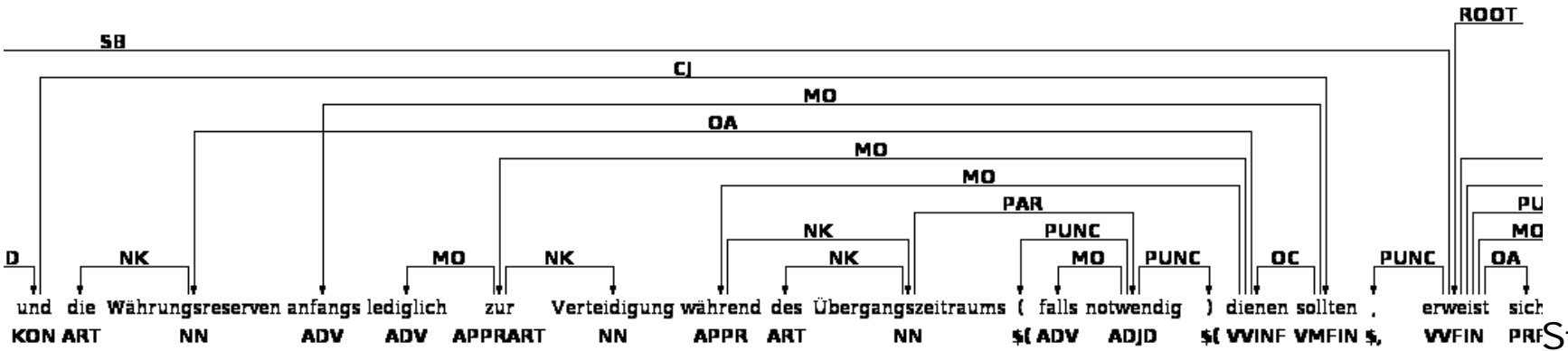
# Reordering automaton example




Starting from **sollten**

- *DOWN*$_{dienen}$ *DOWN*$_{Wahr...}$ *DOWN*$_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**
- *UP UP DOWN*$_{anfangs}$ *EMIT* **anfangs**

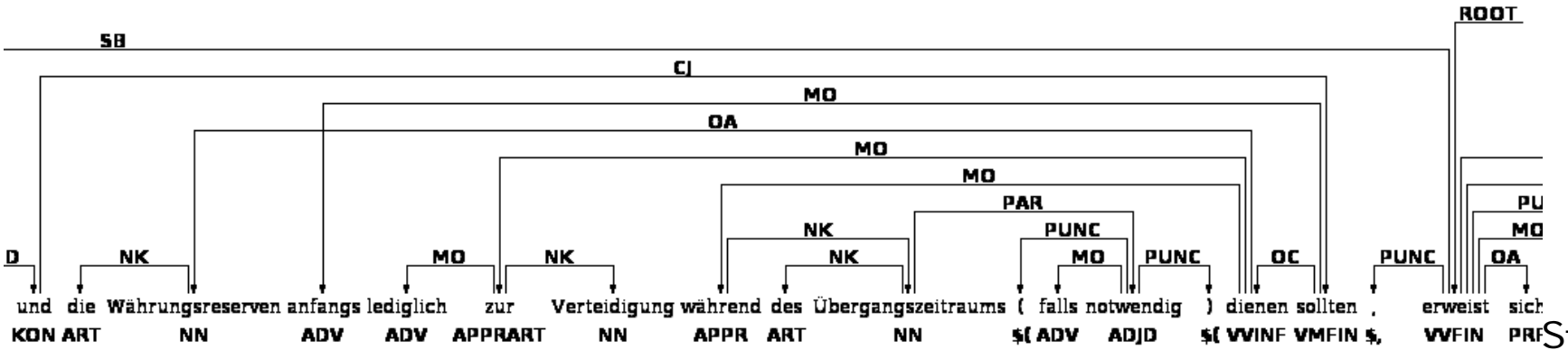
# Reordering automaton example




Starting from **sollten**

- $DOWN_{dienen}$ $DOWN_{Wahr...}$ $DOWN_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**
- *UP UP* $DOWN_{anfangs}$ *EMIT* **anfangs**
- *UP* $DOWN_{dienen}$ $DOWN_{zur}$ $DOWN_{ledi...}$ *EMIT* **lediglich**

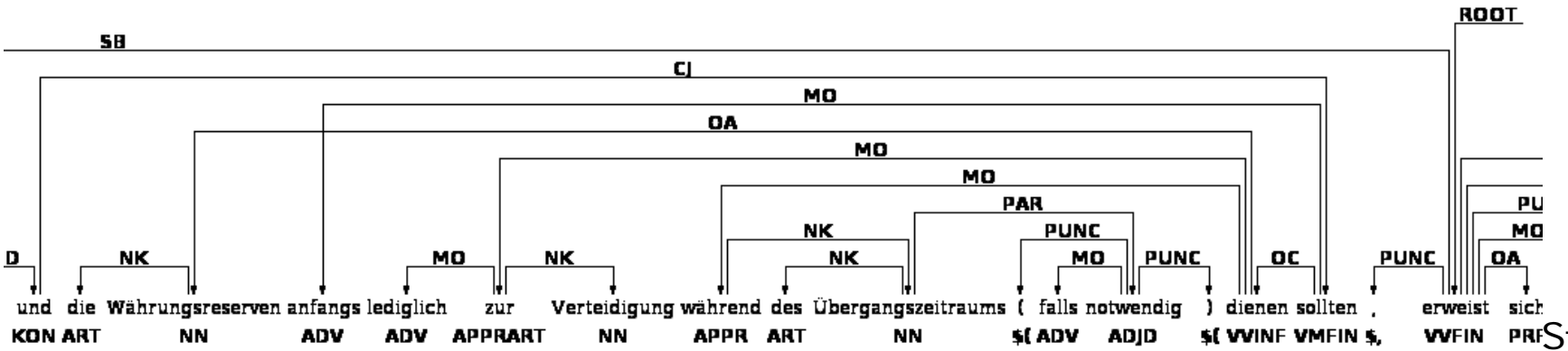
# Reordering automaton example




Starting from **sollten**

- *DOWN*$_{dienen}$ *DOWN*$_{Wahr...}$ *DOWN*$_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**
- *UP UP DOWN*$_{anfangs}$ *EMIT* **anfangs**
- *UP DOWN*$_{dienen}$ *DOWN*$_{zur}$ *DOWN*$_{ledi...}$ *EMIT* **lediglich**
- *UP UP EMIT* **dienen**

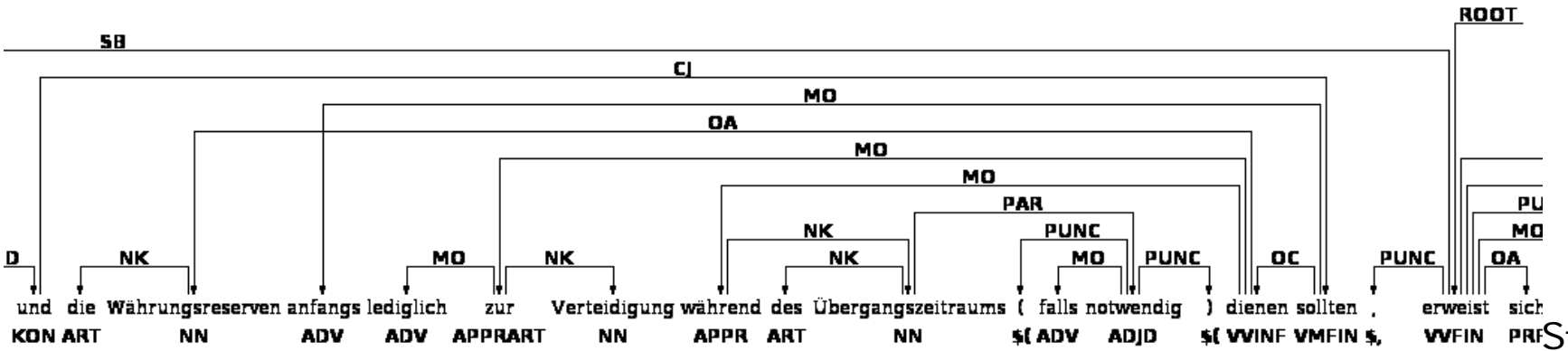
# Reordering automaton example




Starting from **sollten**

- *DOWN*$_{dienen}$ *DOWN*$_{Wahr...}$ *DOWN*$_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**
- *UP UP DOWN*$_{anfangs}$ *EMIT* **anfangs**
- *UP DOWN*$_{dienen}$ *DOWN*$_{zur}$ *DOWN*$_{ledi...}$ *EMIT* **lediglich**
- *UP UP EMIT* **dienen**
- *UP EMIT* **sollten**

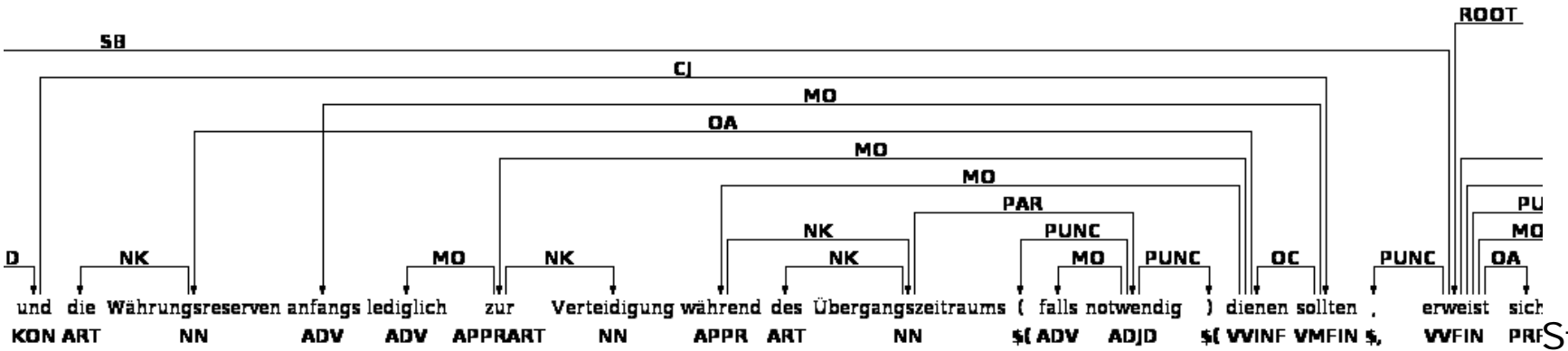
# Reordering automaton example




Starting from **sollten**

- *DOWN*$_{dienen}$ *DOWN*$_{Wahr...}$ *DOWN*$_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**
- *UP UP DOWN*$_{anfangs}$ *EMIT* **anfangs**
- *UP DOWN*$_{dienen}$ *DOWN*$_{zur}$ *DOWN*$_{ledi...}$ *EMIT* **lediglich**
- *UP UP EMIT* **dienen**
- *UP EMIT* **sollten**
- *DOWN*$_{dienen}$ *DOWN*$_{zur}$ *EMIT* **zur**

# Reordering automaton example




Starting from **sollten**

- $DOWN_{dienen}$ $DOWN_{Wahr...}$ $DOWN_{die}$ *EMIT* **die**
- *UP EMIT* **Währungsreserven**
- *UP UP* $DOWN_{anfangs}$ *EMIT* **anfangs**
- *UP* $DOWN_{dienen}$ $DOWN_{zur}$ $DOWN_{ledi...}$ *EMIT* **lediglich**
- *UP UP EMIT* **dienen**
- *UP EMIT* **sollten**
- $DOWN_{dienen}$ $DOWN_{zur}$ *EMIT* **zur**
- $DOWN_{Vert...}$ *EMIT* **Verteidigung**

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?
  - Use a recurrent neural network!

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?
  - Use a recurrent neural network!
- Two-level hierarchical recurrent neural network

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?
  - Use a recurrent neural network!
- Two-level hierarchical recurrent neural network
  - Inner recurrence: one step per action in a fragment

$$v_r(t) = \tau(\Theta^{(r_1)} \cdot x_r(t_r) + \Theta^{r_{REC}} \cdot v_r(t_r - 1)) \qquad (2)$$

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?
  - Use a recurrent neural network!
- Two-level hierarchical recurrent neural network
  - Inner recurrence: one step per action in a fragment

$$v_r(t) = \tau(\Theta^{(r_1)} \cdot x_r(t_r) + \Theta^{r_{REC}} \cdot v_r(t_r - 1)) \qquad (2)$$

  - Outer recurrence: one step per fragment, takes inner state vector as a feature

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?
  - Use a recurrent neural network!
- Two-level hierarchical recurrent neural network
  - Inner recurrence: one step per action in a fragment

$$v_r(t) = \tau(\Theta^{(r_1)} \cdot x_r(t_r) + \Theta^{r_{REC}} \cdot v_r(t_r - 1)) \qquad (2)$$

  - Outer recurrence: one step per fragment, takes inner state vector as a feature
- Everything is differentiable, train end-to-end with generalized backpropagation through time .

# Fragment RNN-RM

- the RNN-RM requires fixed-size per-word features. Fragments have variable size
  - How do we transform fragments into fixed-size representations?
  - Use a recurrent neural network!
- Two-level hierarchical recurrent neural network
  - Inner recurrence: one step per action in a fragment

$$v_r(t) = \tau(\Theta^{(r_1)} \cdot x_r(t_r) + \Theta^{r_{REC}} \cdot v_r(t_r - 1)) \qquad (2)$$

  - Outer recurrence: one step per fragment, takes inner state vector as a feature
- Everything is differentiable, train end-to-end with generalized backpropagation through time . (Theano)

# Results

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |

Translation BLEU:

| Test set | system | BLEU | improvement |
|---|---|---|---|
| Europarl | baseline | 33.00 | |
| Europarl | ″oracle″ | 41.80 | +8.80 |
| Europarl | Collins | 33.52 | +0.52 |
| Europarl | unlex. Base RNN-RM | 33.41 | +0.41 |
| Europarl | lex. Base RNN-RM | 33.38 | +0.38 |
| Europarl | unlex. Fragment RNN-RM | 33.54 | +0.54 |
| news2013 | baseline | 18.80 | |
| news2013 | Collins | NA | NA |
| news2013 | unlex. Base RNN-RM | 19.19 | +0.39 |
| news2013 | lex. Base RNN-RM | 19.01 | +0.21 |
| news2013 | unlex. Fragment RNN-RM | 19.27 | +0.47 |
| news2009 | baseline | 18.09 | |
| news2009 | **Collins** | 18.74 | **+0.65** |
| news2009 | unlex. Base RNN-RM | 18.50 | +0.41 |
| news2009 | lex. Base RNN-RM | 18.44 | +0.35 |
| news2009 | unlex. Fragment RNN-RM | 18.60 | +0.51 |

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |

Translation BLEU:

| Test set | system | BLEU | improvement |
|---|---|---|---|
| Europarl | baseline | 33.00 | |
| Europarl | ″oracle″ | 41.80 | +8.80 |
| Europarl | Collins | 33.52 | +0.52 |
| Europarl | unlex. Base RNN-RM | 33.41 | +0.41 |
| Europarl | lex. Base RNN-RM | 33.38 | +0.38 |
| Europarl | unlex. Fragment RNN-RM | 33.54 | +0.54 |
| news2013 | baseline | 18.80 | |
| news2013 | Collins | NA | NA |
| news2013 | unlex. Base RNN-RM | 19.19 | +0.39 |
| news2013 | lex. Base RNN-RM | 19.01 | +0.21 |
| news2013 | unlex. Fragment RNN-RM | 19.27 | +0.47 |
| news2009 | baseline | 18.09 | |
| news2009 | **Collins** | 18.74 | **+0.65** |
| news2009 | unlex. Base RNN-RM | 18.50 | +0.41 |
| news2009 | lex. Base RNN-RM | 18.44 | +0.35 |
| news2009 | unlex. Fragment RNN-RM | 18.60 | +0.51 |

☹Slow!

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |

Translation BLEU:

| Test set | system | BLEU | improvement |
|---|---|---|---|
| Europarl | baseline | 33.00 | |
| Europarl | ″oracle″ | 41.80 | +8.80 |
| Europarl | Collins | 33.52 | +0.52 |
| Europarl | unlex. Base RNN-RM | 33.41 | +0.41 |
| Europarl | lex. Base RNN-RM | 33.38 | +0.38 |
| Europarl | unlex. Fragment RNN-RM | 33.54 | +0.54 |
| news2013 | baseline | 18.80 | |
| news2013 | Collins | NA | NA |
| news2013 | unlex. Base RNN-RM | 19.19 | +0.39 |
| news2013 | lex. Base RNN-RM | 19.01 | +0.21 |
| news2013 | unlex. Fragment RNN-RM | 19.27 | +0.47 |
| news2009 | baseline | 18.09 | |
| news2009 | **Collins** | 18.74 | **+0.65** |
| news2009 | unlex. Base RNN-RM | 18.50 | +0.41 |
| news2009 | lex. Base RNN-RM | 18.44 | +0.35 |
| news2009 | unlex. Fragment RNN-RM | 18.60 | +0.51 |

☹Slow! $O(L^3)$.

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |

Translation BLEU:

| Test set | system | BLEU | improvement |
|---|---|---|---|
| Europarl | baseline | 33.00 | |
| Europarl | ″oracle″ | 41.80 | +8.80 |
| Europarl | Collins | 33.52 | +0.52 |
| Europarl | unlex. Base RNN-RM | 33.41 | +0.41 |
| Europarl | lex. Base RNN-RM | 33.38 | +0.38 |
| Europarl | unlex. Fragment RNN-RM | 33.54 | +0.54 |
| news2013 | baseline | 18.80 | |
| news2013 | Collins | NA | NA |
| news2013 | unlex. Base RNN-RM | 19.19 | +0.39 |
| news2013 | lex. Base RNN-RM | 19.01 | +0.21 |
| news2013 | unlex. Fragment RNN-RM | 19.27 | +0.47 |
| news2009 | baseline | 18.09 | |
| news2009 | **Collins** | 18.74 | **+0.65** |
| news2009 | unlex. Base RNN-RM | 18.50 | +0.41 |
| news2009 | lex. Base RNN-RM | 18.44 | +0.35 |
| news2009 | unlex. Fragment RNN-RM | 18.60 | +0.51 |

☹Slow! $O(L^3)$. 5 days of training, 3 days decoding (no GPU)

# Gated Recurrent Unit Reordering Model

- Standard (Elman) recurrent neural networks are hard to train and don't capture well long-distance dependencies

# Gated Recurrent Unit Reordering Model

- Standard (Elman) recurrent neural networks are hard to train and don't capture well long-distance dependencies
- Gated Recurrent Unit [Cho et al. 2014]

# Gated Recurrent Unit Reordering Model

- Standard (Elman) recurrent neural networks are hard to train and don't capture well long-distance dependencies
- Gated Recurrent Unit [Cho et al. 2014]
- similar to LSTM

# Gated Recurrent Unit Reordering Model

- Standard (Elman) recurrent neural networks are hard to train and don't capture well long-distance dependencies
- Gated Recurrent Unit [Cho et al. 2014]
- similar to LSTM

$$
\begin{aligned}
v_{rst}(t) &= \pi(\Theta^{(1)}_{rst} \cdot x(t) + \Theta^{REC}_{rst} \cdot v(t-1)) \\
v_{upd}(t) &= \pi(\Theta^{(1)}_{upd} \cdot x(t) + \Theta^{REC}_{upd} \cdot v(t-1)) \\
v_{raw}(t) &= \tau(\Theta^{(1)} \cdot x(t) + \Theta^{REC} \cdot v(t-1) \odot v_{upd}(t)) \\
v(t) &= v_{rst}(t) \odot v(t-1) + (1 - v_{rst}(t)) \odot v_{raw}(t)
\end{aligned}
\tag{3}
$$

# Results

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |
| unlex. Base GRU-RM | 64.78 | +2.68 |

# Results

Monolingual BLEU:

| Reordering | BLEU | improvement |
|---|---|---|
| none | 62.10 | |
| unlex. Base RNN-RM | 64.03 | +1.93 |
| lex. Base RNN-RM | 63.99 | +1.89 |
| unlex. Fragment RNN-RM | 64.43 | +2.33 |
| unlex. Base GRU-RM | 64.78 | +2.68 |

Translation BLEU:

| Test set | system | BLEU | improvement |
|---|---|---|---|
| Europarl | baseline | 33.00 | |
| Europarl | ”oracle” | 41.80 | +8.80 |
| Europarl | Collins | 33.52 | +0.52 |
| Europarl | unlex. Base RNN-RM | 33.41 | +0.41 |
| Europarl | lex. Base RNN-RM | 33.38 | +0.38 |
| Europarl | unlex. Fragment RNN-RM | 33.54 | +0.54 |
| Europarl | **unlex. Base GRU-RM** | 34.15 | **+1.15** |
| news2013 | baseline | 18.80 | |
| news2013 | Collins | NA | NA |
| news2013 | unlex. Base RNN-RM | 19.19 | +0.39 |
| news2013 | lex. Base RNN-RM | 19.01 | +0.21 |
| news2013 | unlex. Fragment RNN-RM | 19.27 | +0.47 |
| news2013 | **unlex. Base GRU-RM** | 19.28 | **+0.48** |
| news2009 | baseline | 18.09 | |
| news2009 | **Collins** | 18.74 | **+0.65** |
| news2009 | unlex. Base RNN-RM | 18.50 | +0.41 |
| news2009 | lex. Base RNN-RM | 18.44 | +0.35 |
| news2009 | unlex. Fragment RNN-RM | 18.60 | +0.51 |

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?
  - Fragment RNN-RM shows that complex moves add value

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?
  - Fragment RNN-RM shows that complex moves add value
  - But it is perhaps too slow for practical applications

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?
  - Fragment RNN-RM shows that complex moves add value
  - But it is perhaps too slow for practical applications
- Base GRU-RM considers less sophisticated features but it is good enough for practical applications

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?
  - Fragment RNN-RM shows that complex moves add value
  - But it is perhaps too slow for practical applications
- Base GRU-RM considers less sophisticated features but it is good enough for practical applications
- Future work: other language pairs

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?
  - Fragment RNN-RM shows that complex moves add value
  - But it is perhaps too slow for practical applications
- Base GRU-RM considers less sophisticated features but it is good enough for practical applications
- Future work: other language pairs

Thanks for your attention

# Conclusions

- Pre-reordering with non-projective dependency syntax and recurrent neural network improves machine translation
- Research question: Do complex moves over a non-projective dependency tree add value or do simple sibling swaps or sequence prediction suffice?
  - Fragment RNN-RM shows that complex moves add value
  - But it is perhaps too slow for practical applications
- Base GRU-RM considers less sophisticated features but it is good enough for practical applications
- Future work: other language pairs

Thanks for your attention
Questions?