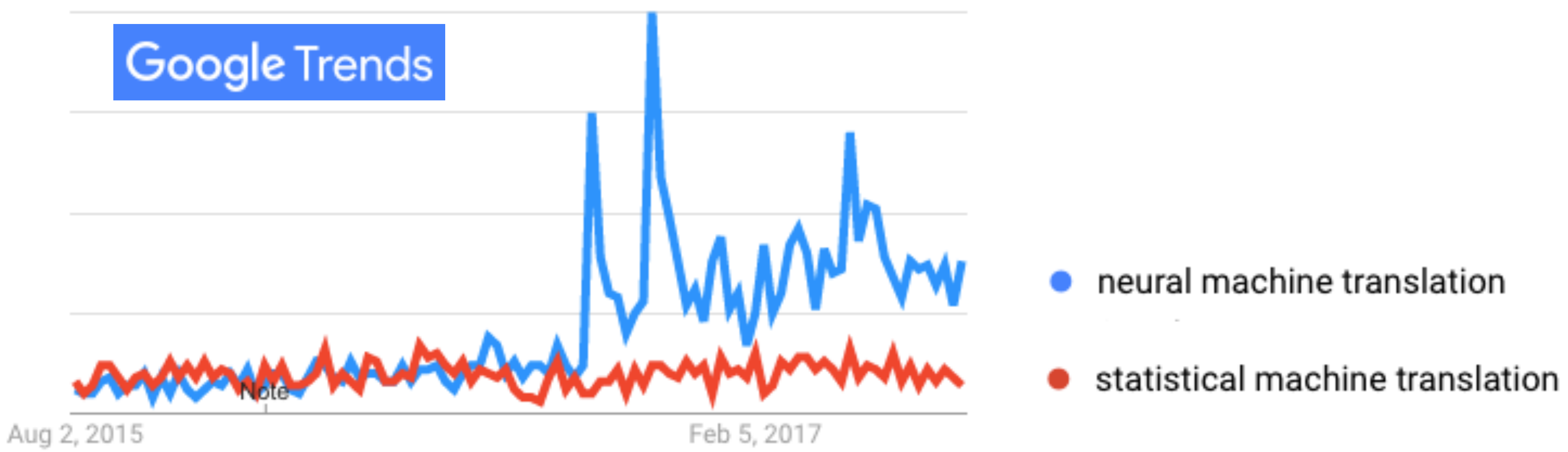
Towards String-to-Tree Neural Machine Translation

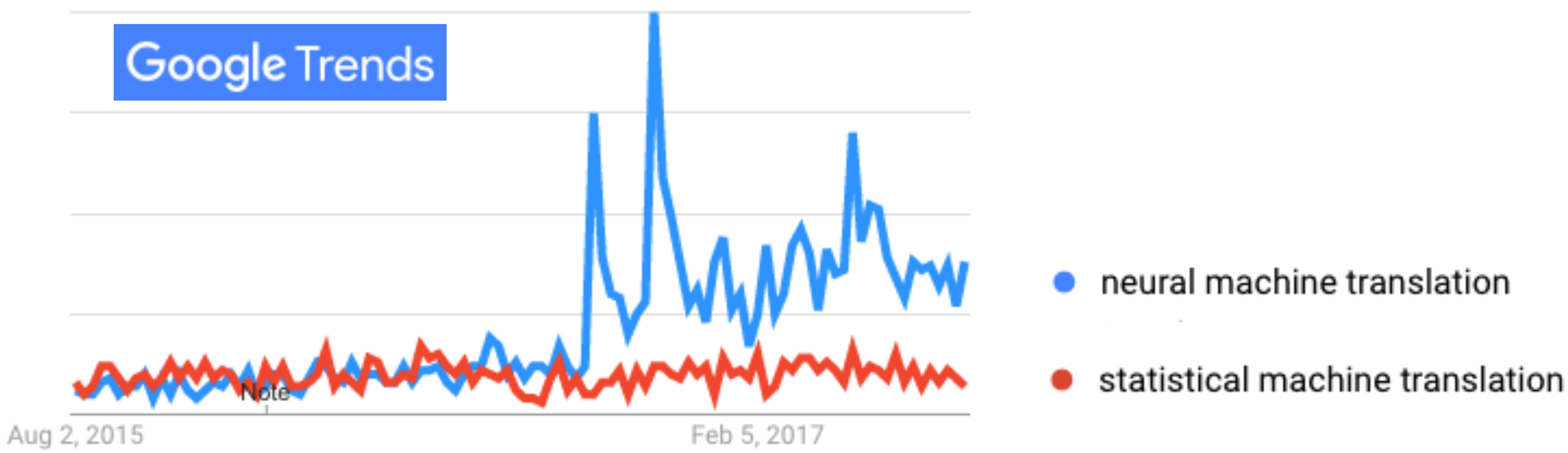
Roee Aharoni & Yoav Goldberg
NLP Lab, Bar Ilan University
ACL 2017

NMT is all the rage!

NMT is all the rage!

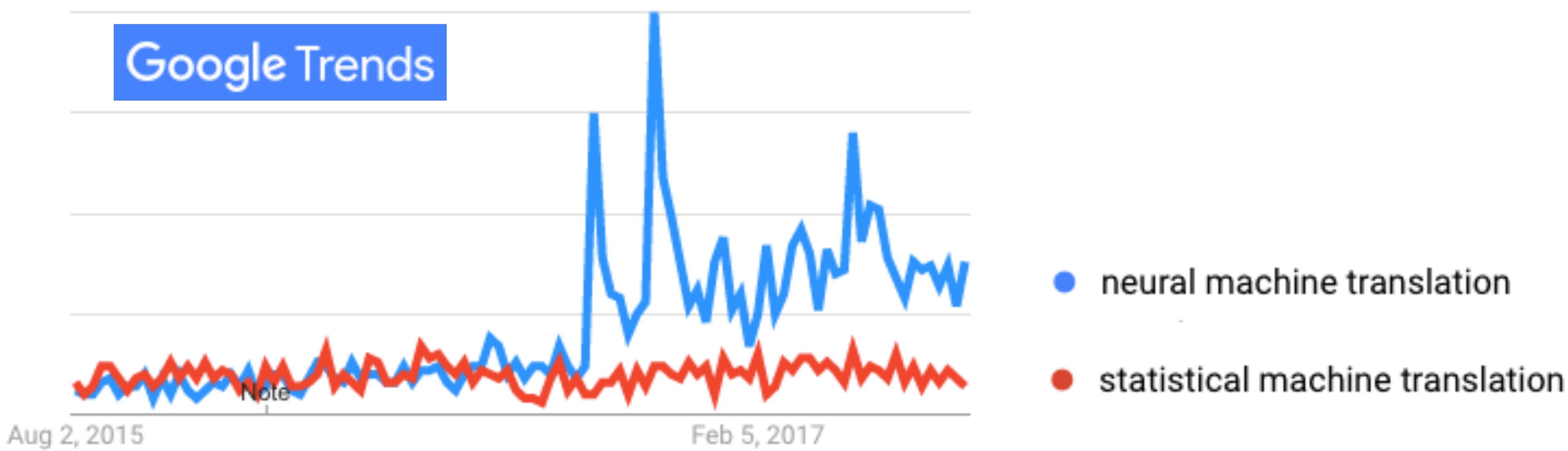


NMT is all the rage!



- Driving the current state-of-the-art (Sennrich et al., 2016)

NMT is all the rage!



- Driving the current state-of-the-art (Sennrich et al., 2016)
- Widely adopted by the industry

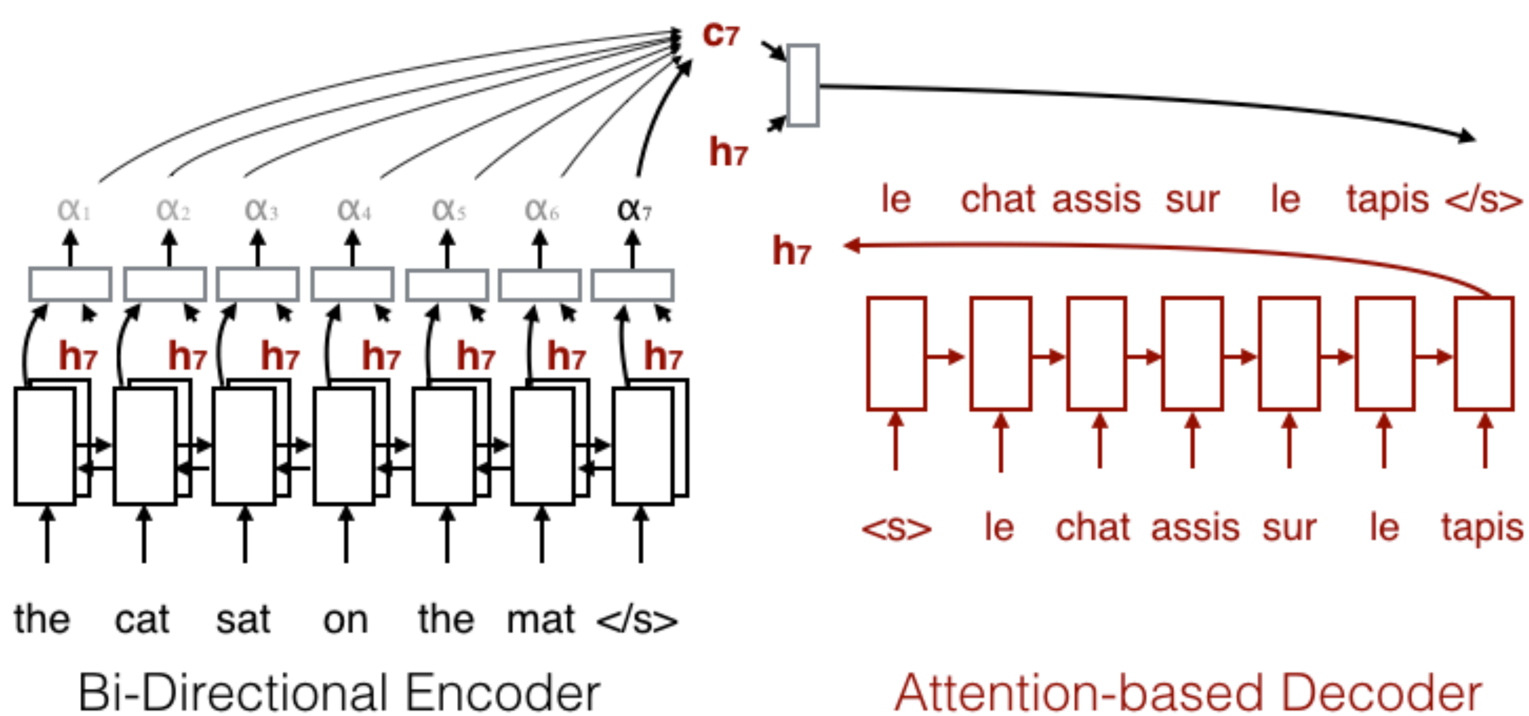
Seq2Seq with Attention

Bahdanau et al. (2015)

Seq2Seq with Attention

Bahdanau et al. (2015)

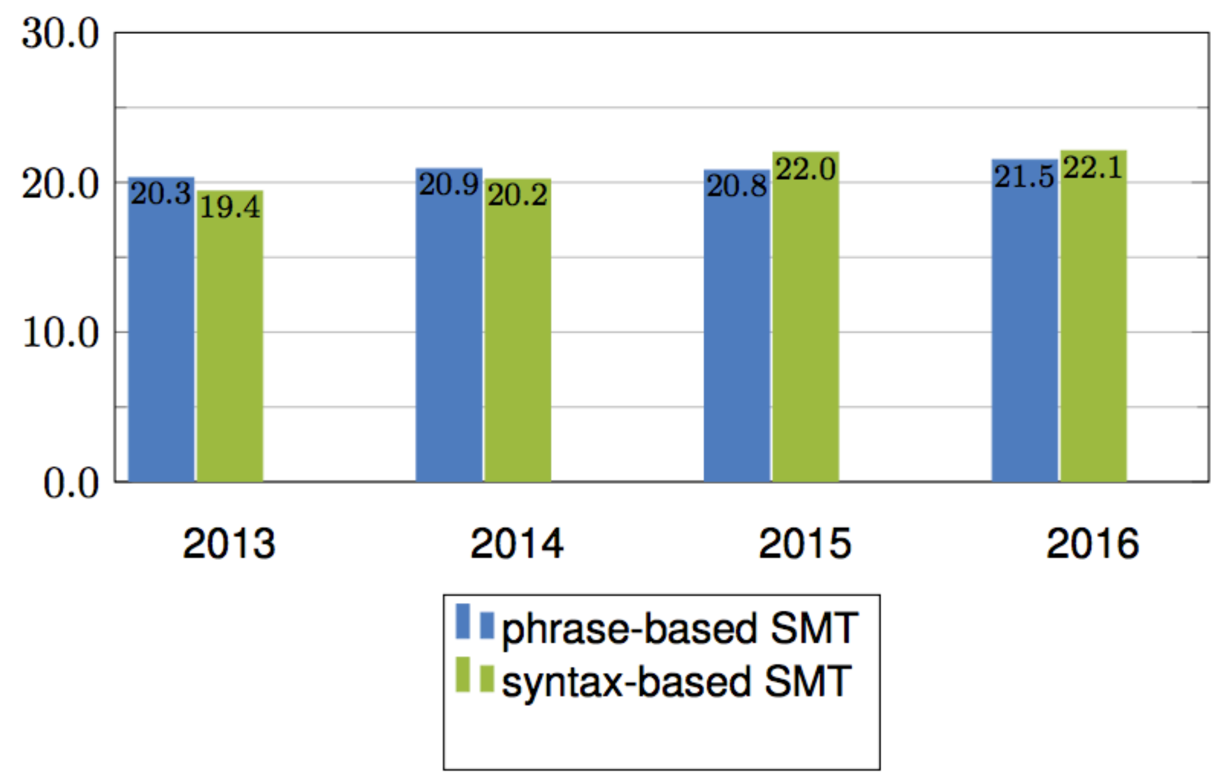
$$f = \operatorname*{argmax}_{f'} p(f'|e)$$



Syntax *was* all the rage!

Syntax *was* all the rage!

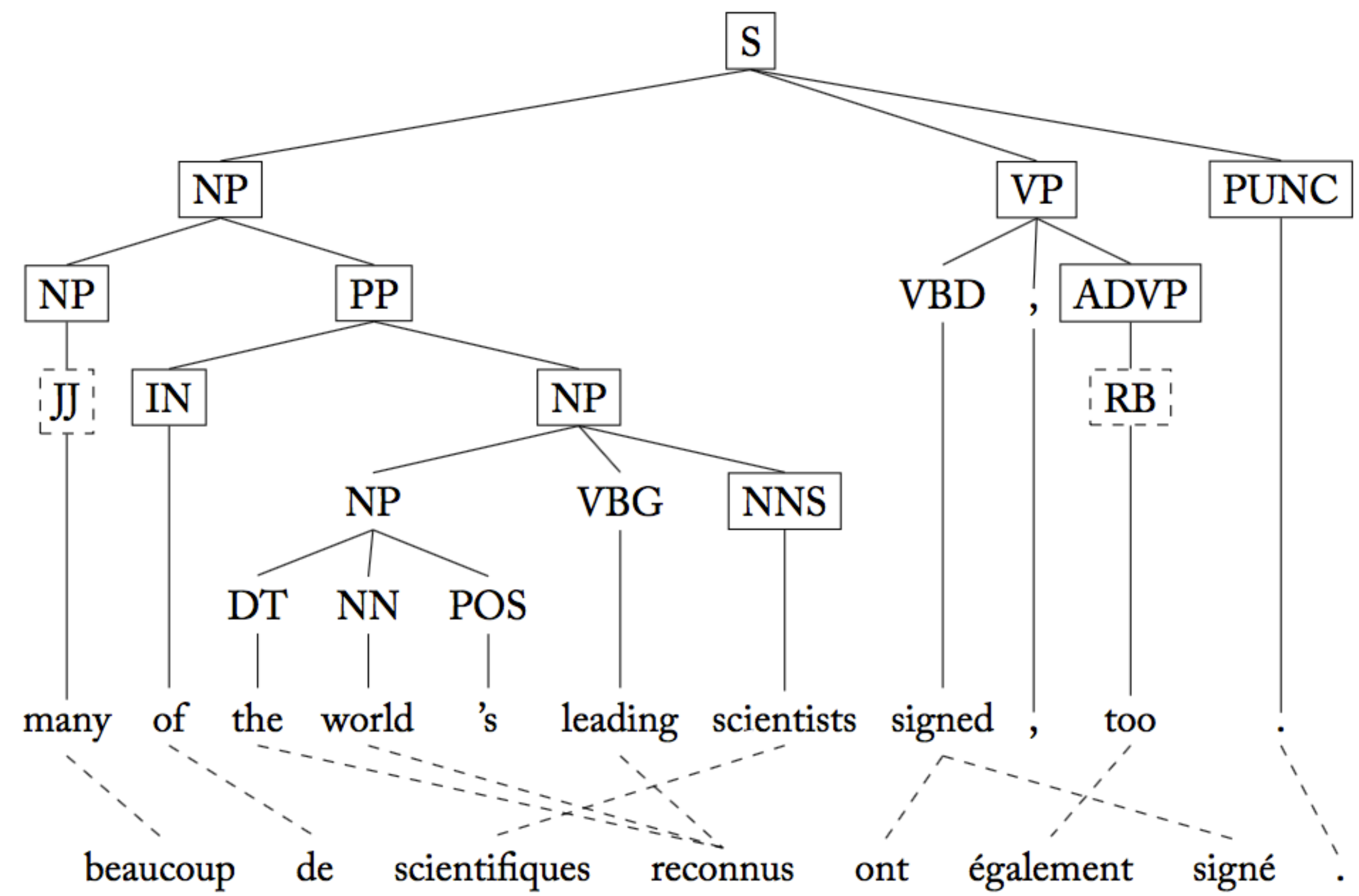
- The "previous" state-of-the-art was syntax-based SMT



From Rico Sennrich, "NMT: Breaking the Performance Plateau", 2016

Syntax *was* all the rage!

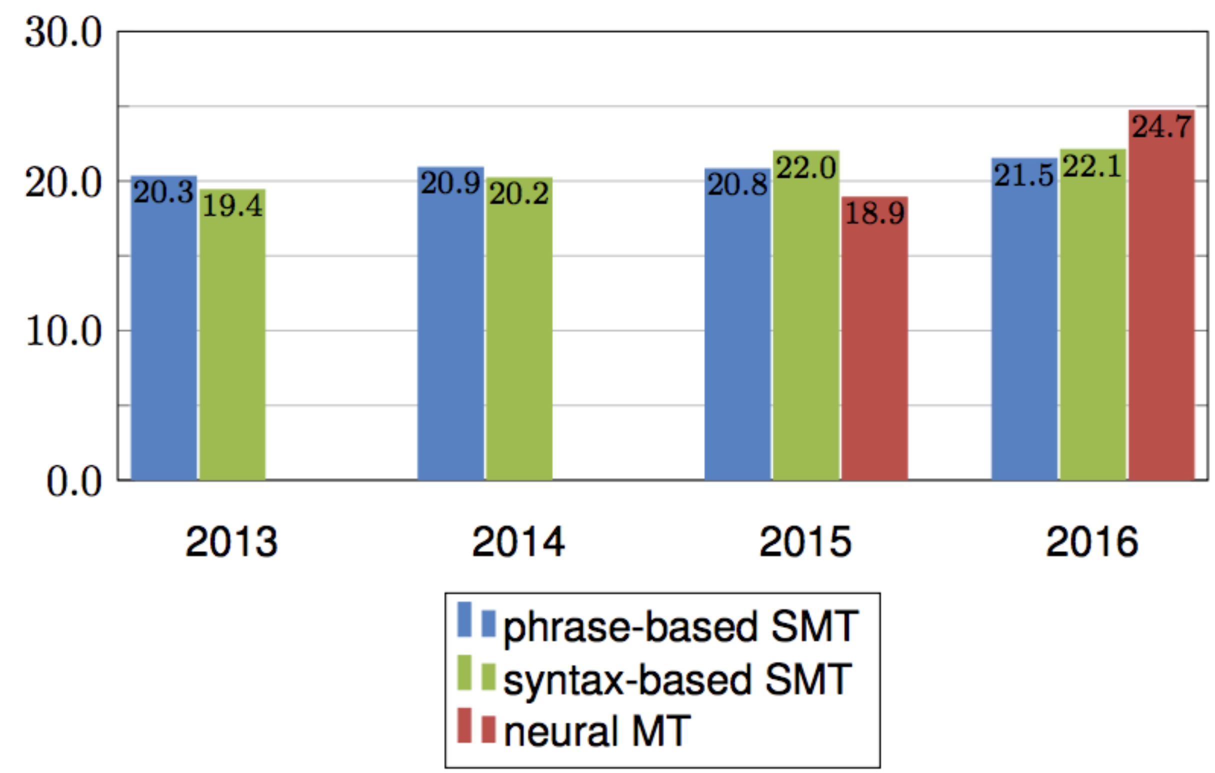
- The “previous” state-of-the-art was syntax-based SMT
- i.e. systems that used linguistic information (usually represented as **parse trees**)



From Williams, Sennrich, Post & Koehn (2016), “Syntax-based Statistical Machine Translation”

Syntax *was* all the rage!

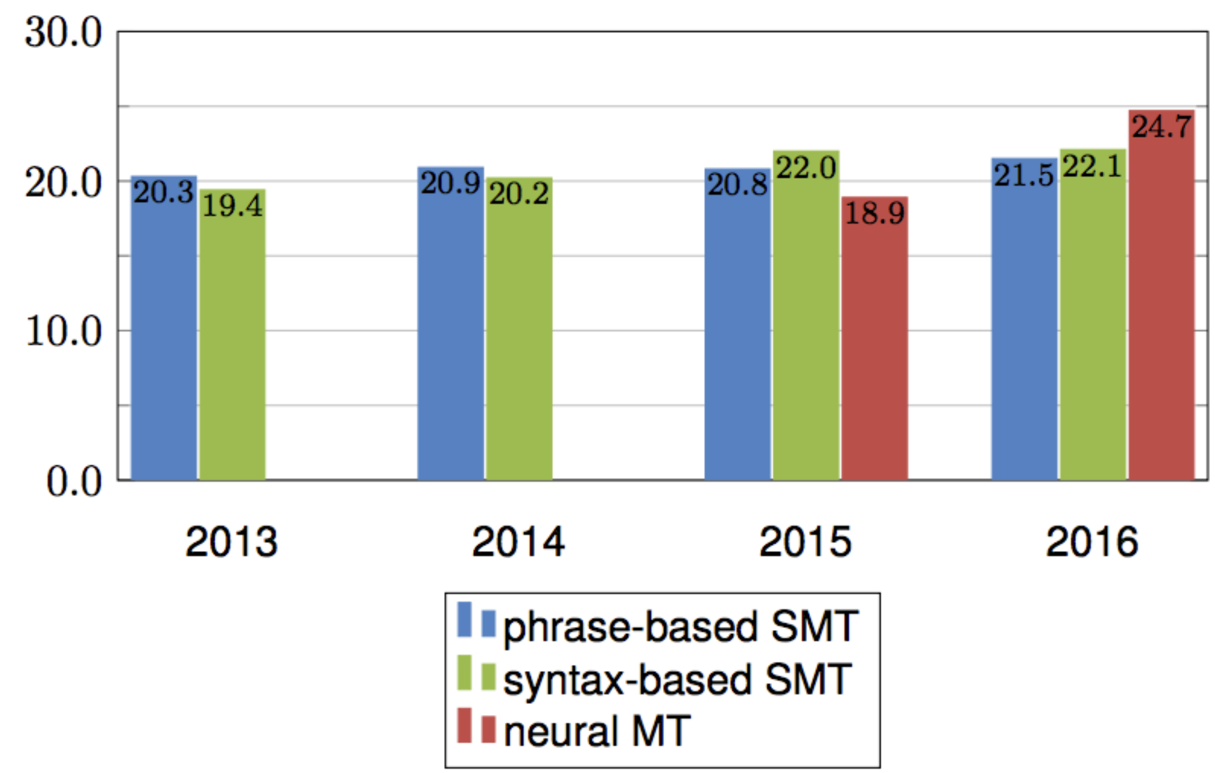
- The "previous" state-of-the-art was syntax-based SMT
- i.e. systems that used linguistic information (usually represented as **parse trees**)
- "Beaten" by NMT in 2016



From Rico Sennrich, "NMT: Breaking the Performance Plateau", 2016

Syntax *was* all the rage!

- The "previous" state-of-the-art was syntax-based SMT
- i.e. systems that used linguistic information (usually represented as **parse trees**)
- "Beaten" by NMT in 2016
- **Can we bring the benefits of syntax into the recent neural systems?**

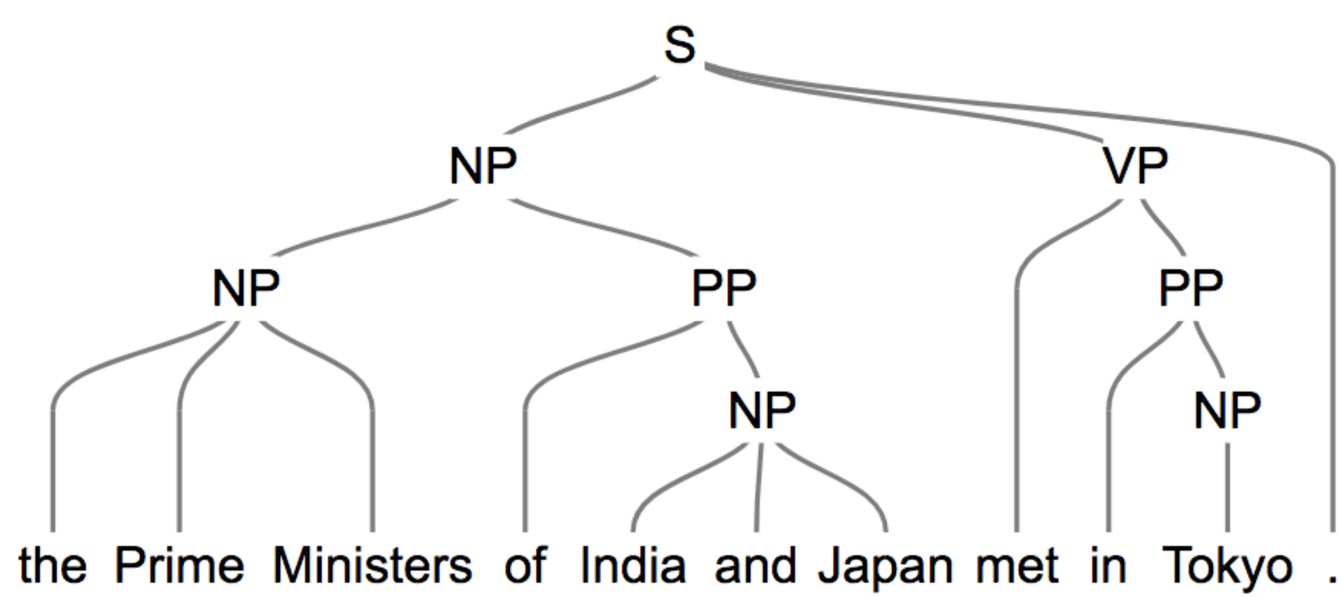


From Rico Sennrich, "NMT: Breaking the Performance Plateau", 2016

Syntax: Constituency Structure

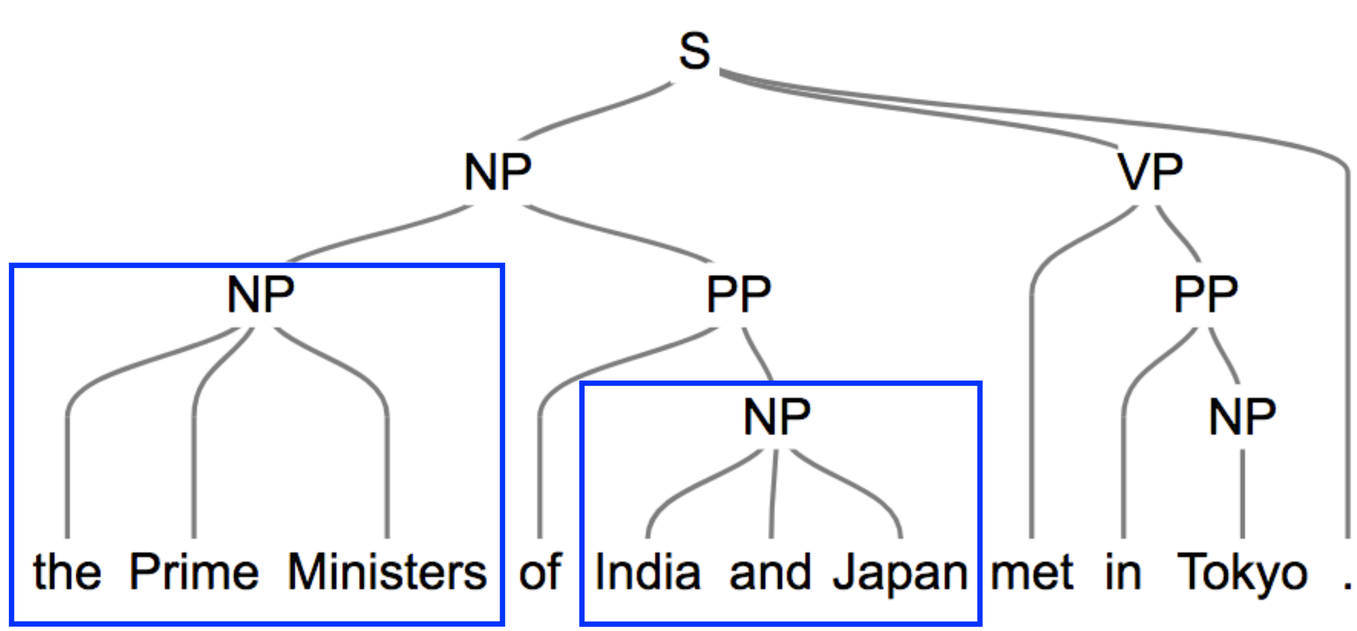
Syntax: Constituency Structure

- A Constituency (a.k.a Phrase-Structure) grammar defines a set of rewrite rules which describe the **structure** of the language.



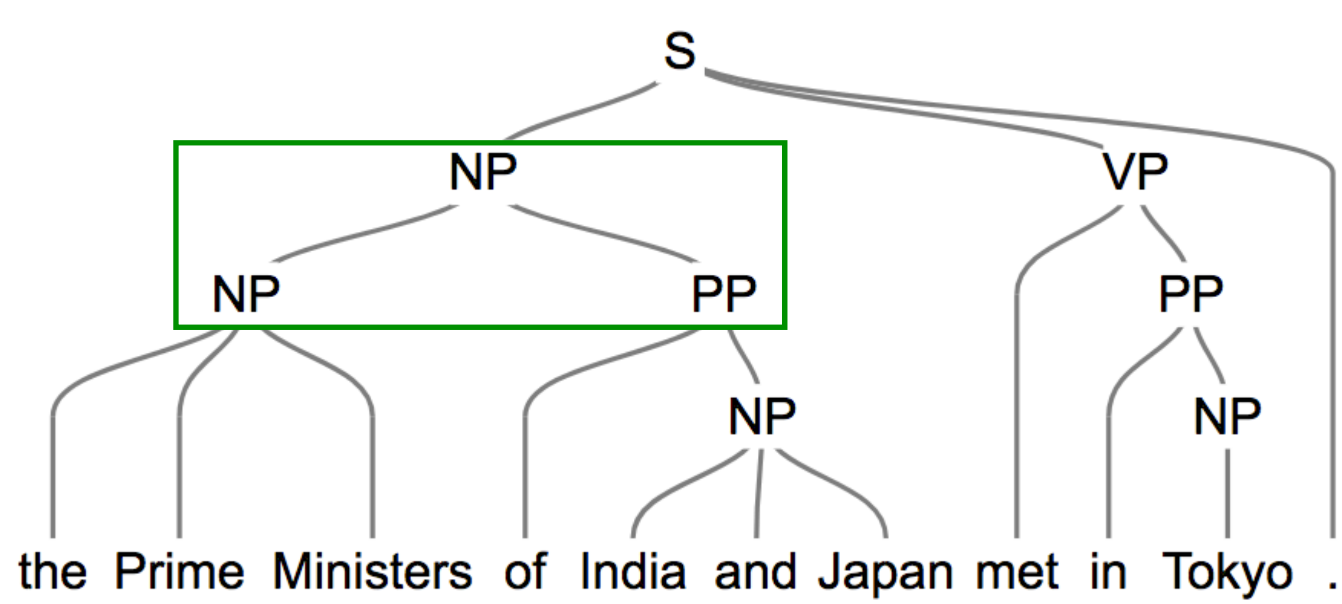
Syntax: Constituency Structure

- A Constituency (a.k.a Phrase-Structure) grammar defines a set of rewrite rules which describe the **structure** of the language.
  - **Groups** words into larger units (constituents)



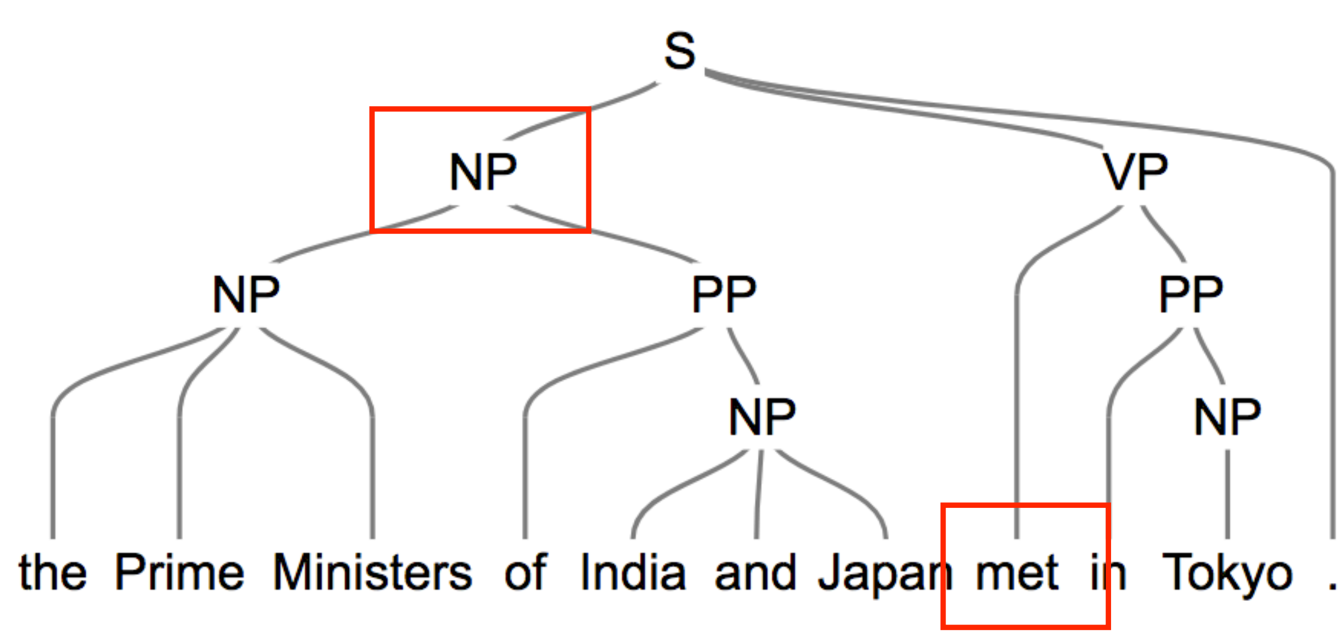
Syntax: Constituency Structure

- A Constituency (a.k.a Phrase-Structure) grammar defines a set of rewrite rules which describe the **structure** of the language.
    - **Groups** words into larger units (constituents)
    - Defines a **hierarchy** between constituents



Syntax: Constituency Structure

- A Constituency (a.k.a Phrase-Structure) grammar defines a set of rewrite rules which describe the **structure** of the language.
  - **Groups** words into larger units (constituents)
  - Defines a **hierarchy** between constituents
  - Draws **relations** between different constituents (words, phrases, clauses…)



Why Syntax Can Help MT?

Why Syntax Can Help MT?

- **Hints** as to which word sequences belong together

Why Syntax Can Help MT?

- **Hints** as to which word sequences belong together
- Helps in producing **well structured** sentences

Why Syntax Can Help MT?

- **Hints** as to which word sequences belong together
- Helps in producing **well structured** sentences
- Allows **informed reordering** decisions according to the syntactic structure

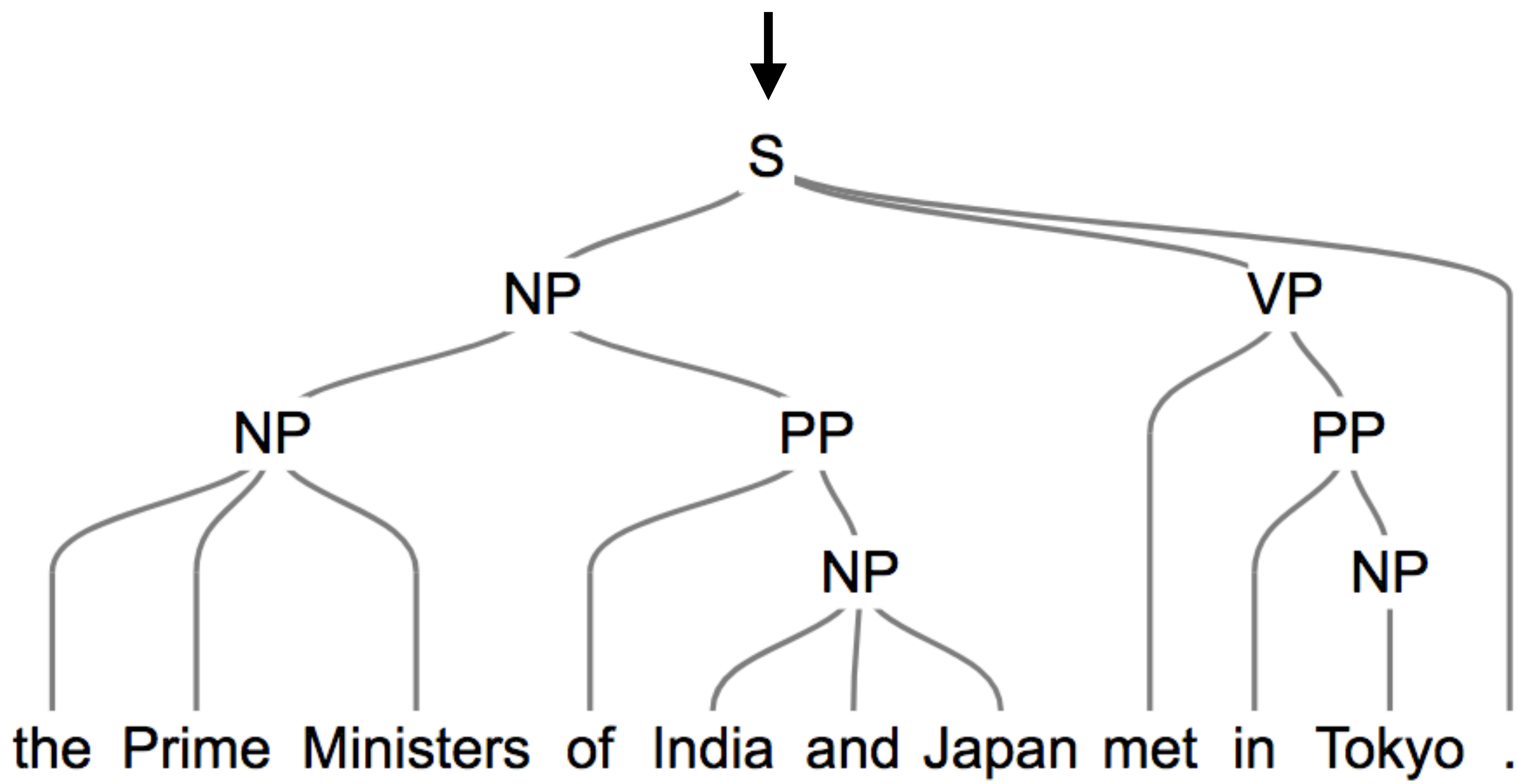
Why Syntax Can Help MT?

- **Hints** as to which word sequences belong together
- Helps in producing **well structured** sentences
- Allows **informed reordering** decisions according to the syntactic structure
- Encourages **long-distance dependencies** when selecting translations

String-to-Tree Translation

source

die Premierminister Indiens und Japans trafen sich in Tokio .



target

Our Approach: String-to-Tree NMT

Our Approach: String-to-Tree NMT

Jane hatte eine Katze .

source

Our Approach: String-to-Tree NMT

Jane hatte eine Katze . $\rightarrow$ **Jane had a cat .**

source **target**

Our Approach: String-to-Tree NMT

Jane hatte eine Katze . $\rightarrow$ $(_{ROOT}$ $(_{S}$ $(_{NP}$ **Jane** $)_{NP}$ $(_{VP}$ **had** $(_{NP}$ **a cat** $)_{NP}$ $)_{VP}$ **.**

source **target**

- Main idea: translate a source sentence into a **linearized tree** of the target sentence

Our Approach: String-to-Tree NMT

Jane hatte eine Katze . $\rightarrow$ $(_{ROOT}$ $(_S$ $(_{NP}$ **Jane** $)_{NP}$ $(_{VP}$ **had** $(_{NP}$ **a cat** $)_{NP}$ $)_{VP}$ **.**

source **target**

- Main idea: translate a source sentence into a **linearized tree** of the target sentence
    - Inspired by works on RNN-based syntactic parsing (Vinyals et. al, 2015, Choe & Charniak, 2016)

Our Approach: String-to-Tree NMT

Jane hatte eine Katze . $\rightarrow$ $(_{ROOT}$ $(_S$ $(_{NP}$ **Jane** $)_{NP}$ $(_{VP}$ **had** $(_{NP}$ **a cat** $)_{NP}$ $)_{VP}$ **.**

source **target**

- Main idea: translate a source sentence into a **linearized tree** of the target sentence
  - Inspired by works on RNN-based syntactic parsing (Vinyals et. al, 2015, Choe & Charniak, 2016)
- Allows using the seq2seq framework as-is

Experimental Details

- We used the Nematus toolkit (Sennrich et al. 2017)
- Joint BPE segmentation (Sennrich et al. 2016)
- For training, we parse the target side using the BLLIP parser (McClosky, Charniak and Johnson, 2006)
- Requires some care about making BPE, Tokenization and Parser work together

Experiments - Large Scale

Experiments - Large Scale

- German to English, **4.5 million** parallel training sentences from WMT16

Experiments - Large Scale

- German to English, **4.5 million** parallel training sentences from WMT16
- Train two NMT models using the same setup (same settings as the SOTA neural system in WMT16)

Experiments - Large Scale

- German to English, **4.5 million** parallel training sentences from WMT16
- Train two NMT models using the same setup (same settings as the SOTA neural system in WMT16)
  - syntax-aware **(bpe2tree)**

Experiments - Large Scale

- German to English, **4.5 million** parallel training sentences from WMT16
- Train two NMT models using the same setup (same settings as the SOTA neural system in WMT16)
  - syntax-aware **(bpe2tree)**
  - syntax-agnostic baseline **(bpe2bpe)**

Experiments - Large Scale

- German to English, **4.5 million** parallel training sentences from WMT16
- Train two NMT models using the same setup (same settings as the SOTA neural system in WMT16)
  - syntax-aware **(bpe2tree)**
  - syntax-agnostic baseline **(bpe2bpe)**
- The syntax-aware model performs better in terms of BLEU

	system	newstest2015	newstest2016
Single Model	bpe2bpe	27.33	31.19
	bpe2tree	27.36	32.13 ←
5 Model Ensemble	bpe2bpe ens.	28.62	32.38
	bpe2tree ens.	28.7	33.24 ←

Experiments - Low Resource

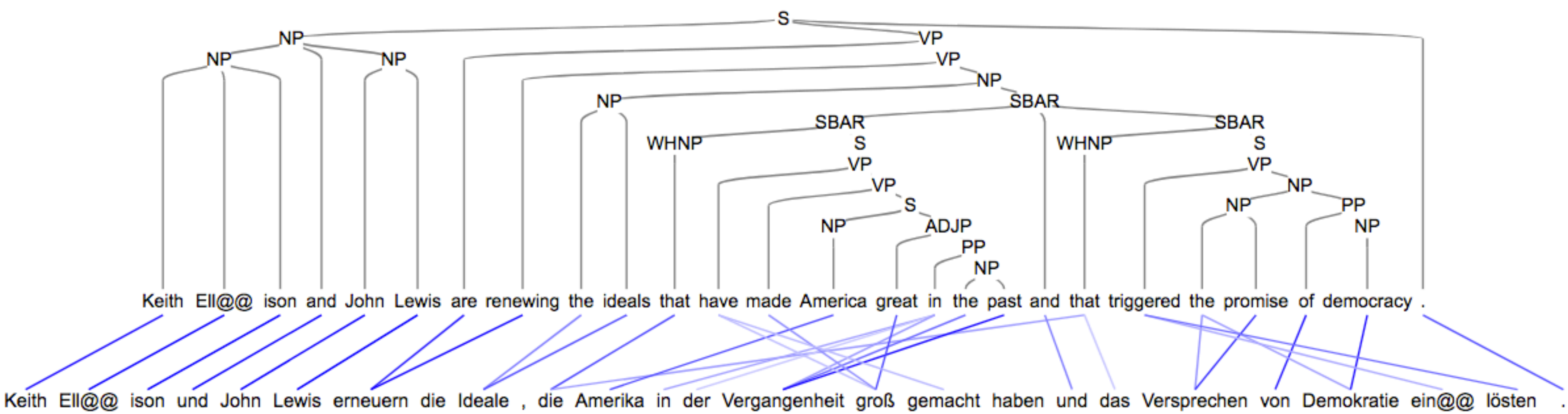
- German/Russian/Czech to English - **180k-140k** parallel training sentences (News Commentary v8)
- The syntax-aware model performs better in terms of BLEU in **all** cases (12 comparisons)
- Up to 2+ BLEU improvement

	system	newstest2015	newstest2016
DE-EN	bpe2bpe	13.81	14.16
	bpe2tree	14.55	16.13
	bpe2bpe ens.	14.42	15.07
	bpe2tree ens.	15.69	17.21
RU-EN	bpe2bpe	12.58	11.37
	bpe2tree	12.92	11.94
	bpe2bpe ens.	13.36	11.91
	bpe2tree ens.	13.66	12.89
CS-EN	bpe2bpe	10.85	11.23
	bpe2tree	11.54	11.65
	bpe2bpe ens.	11.46	11.77
	bpe2tree ens.	12.43	12.68

Looking Beyond BLEU

Accurate Trees

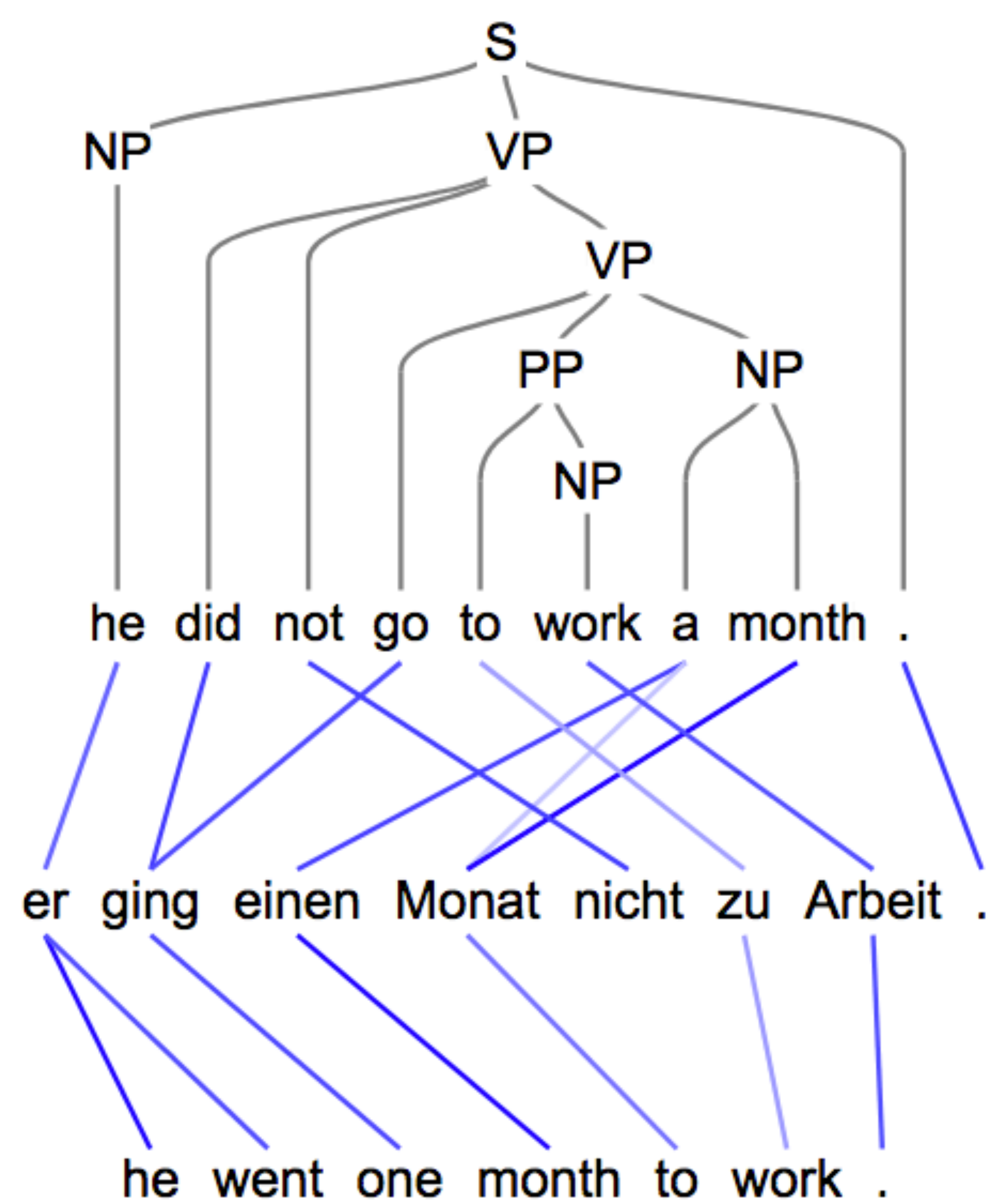
- 99% of the predicted trees in the development set had valid bracketing
- Eye-balling the predicted trees found them well-formed and following the syntax of English.



Where Syntax Helps? Alignments

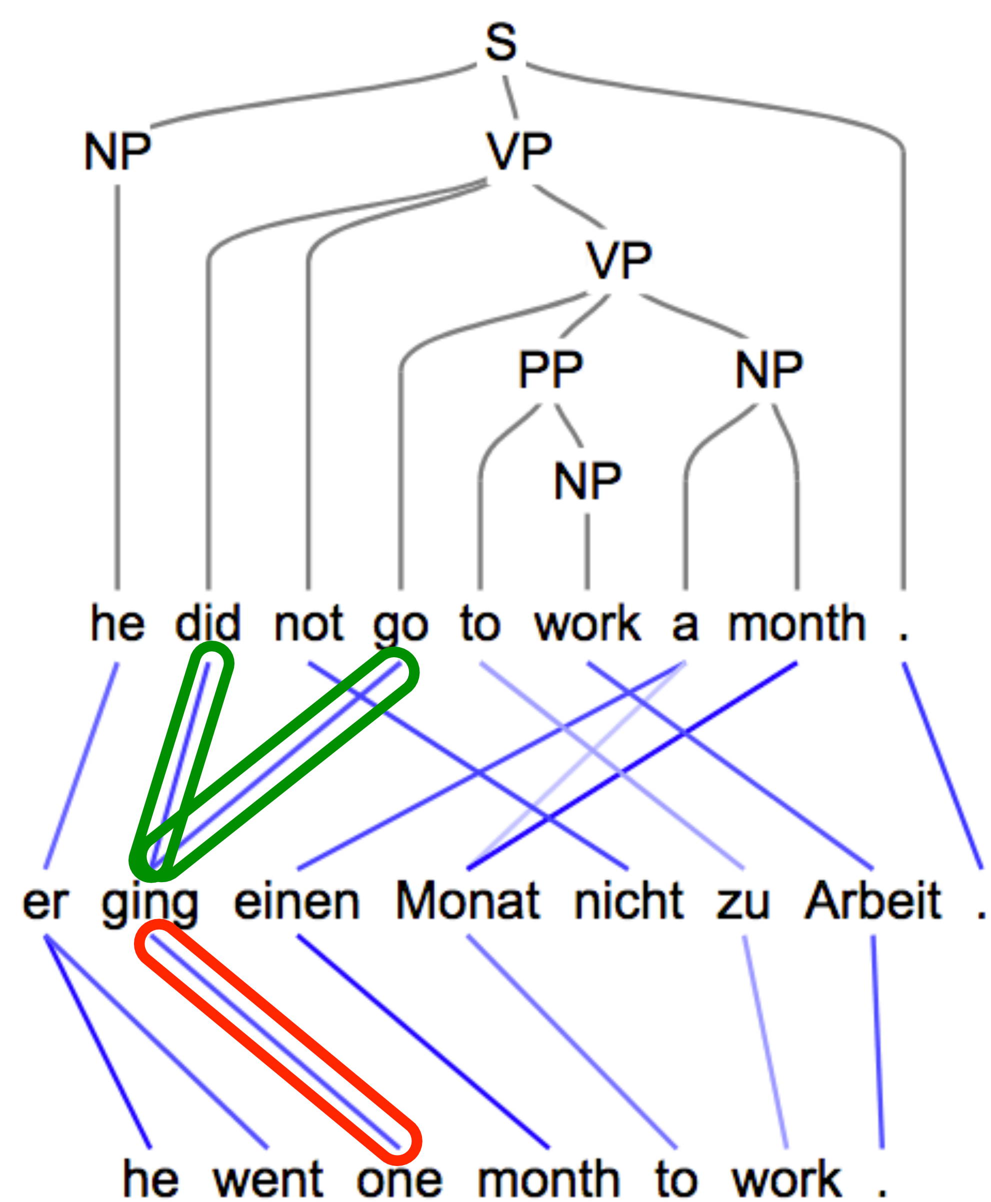
Where Syntax Helps? Alignments

- The attention based model induces soft **alignments** between the source and the target



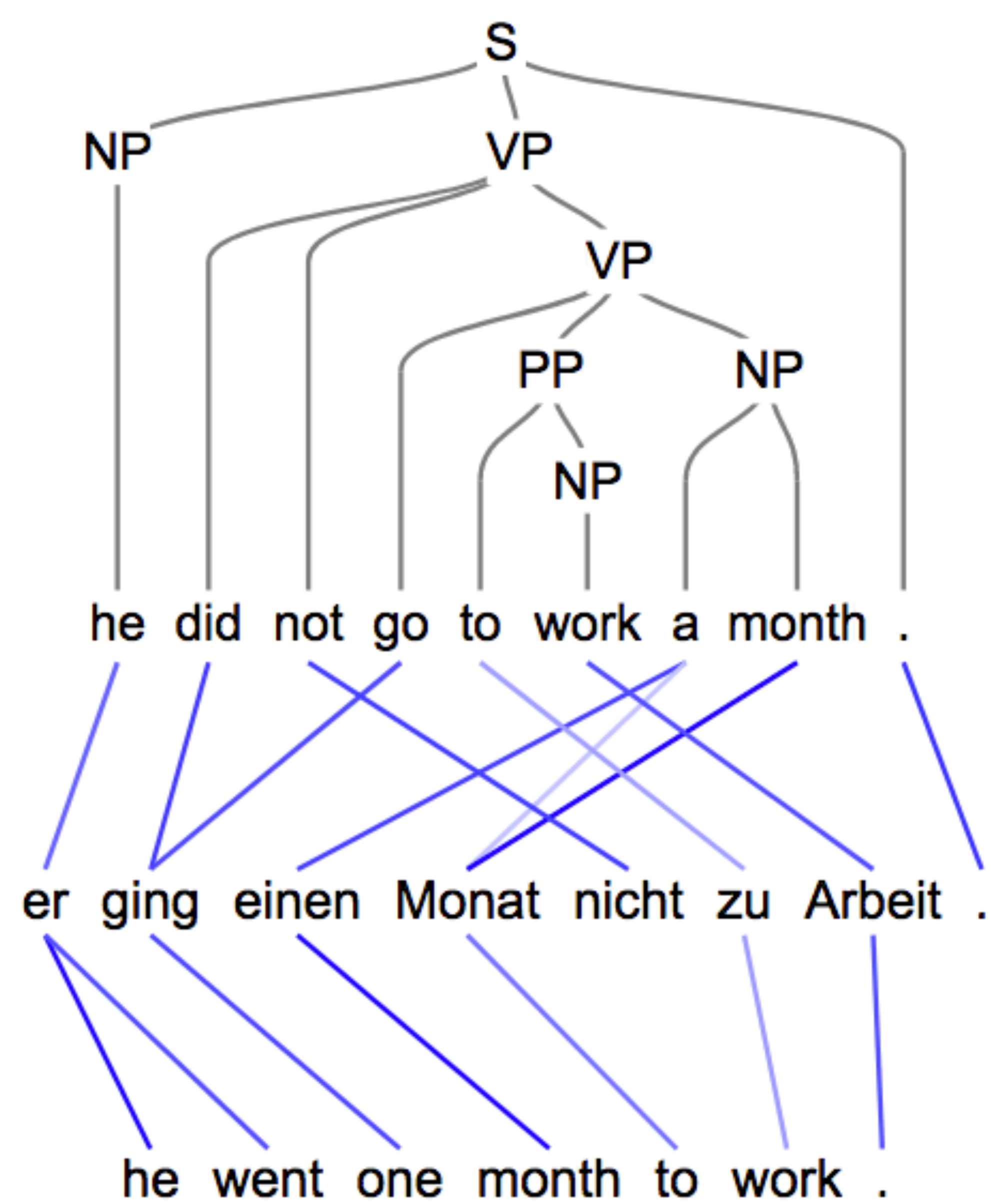
Where Syntax Helps? Alignments

- The attention based model induces soft **alignments** between the source and the target
- The syntax-aware model produced more sensible alignments



Where Syntax Helps? Alignments

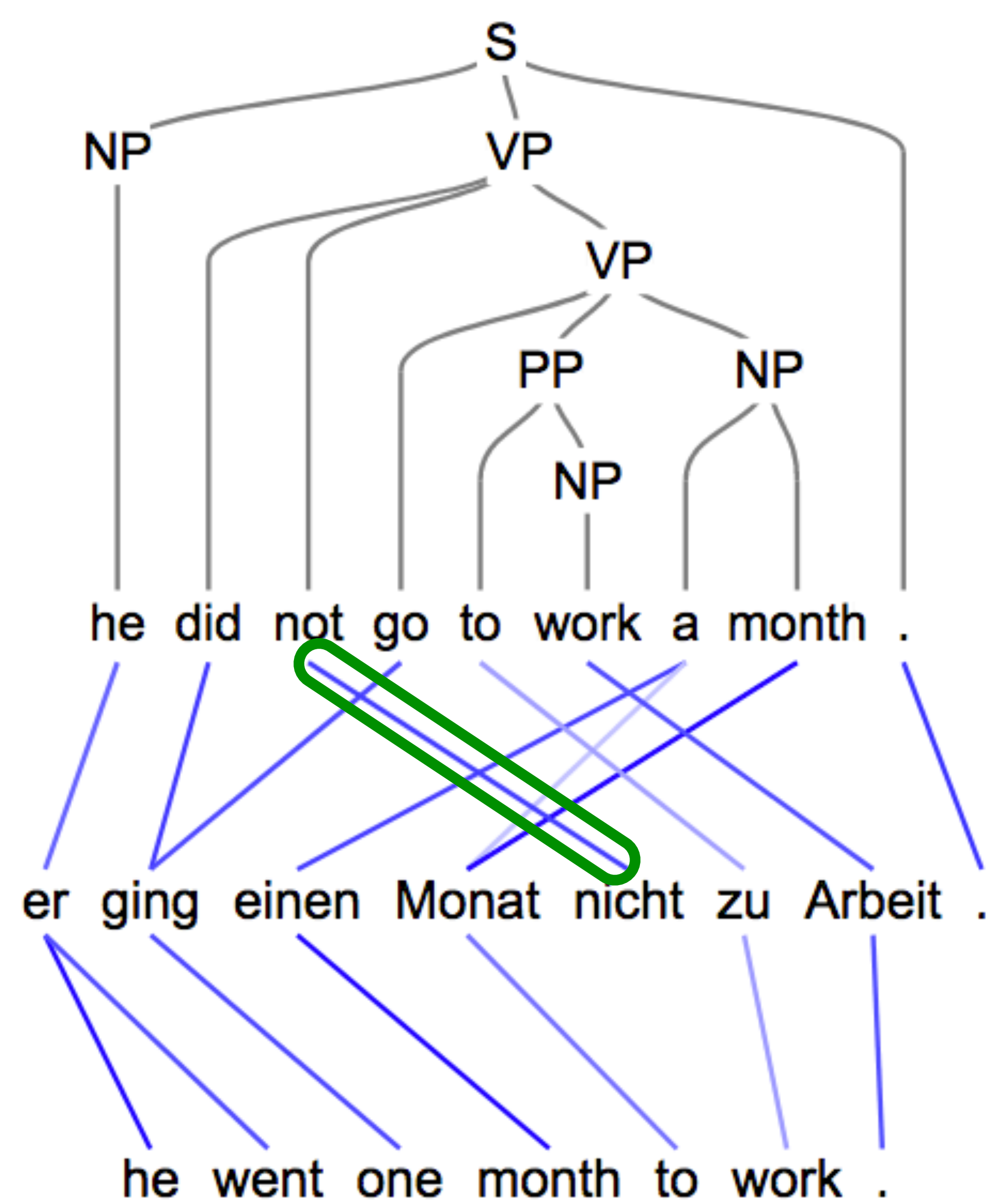
- The attention based model induces soft **alignments** between the source and the target
- The syntax-aware model produced more sensible alignments



Where Syntax Helps? Alignments

- The attention based model induces soft **alignments** between the source and the target

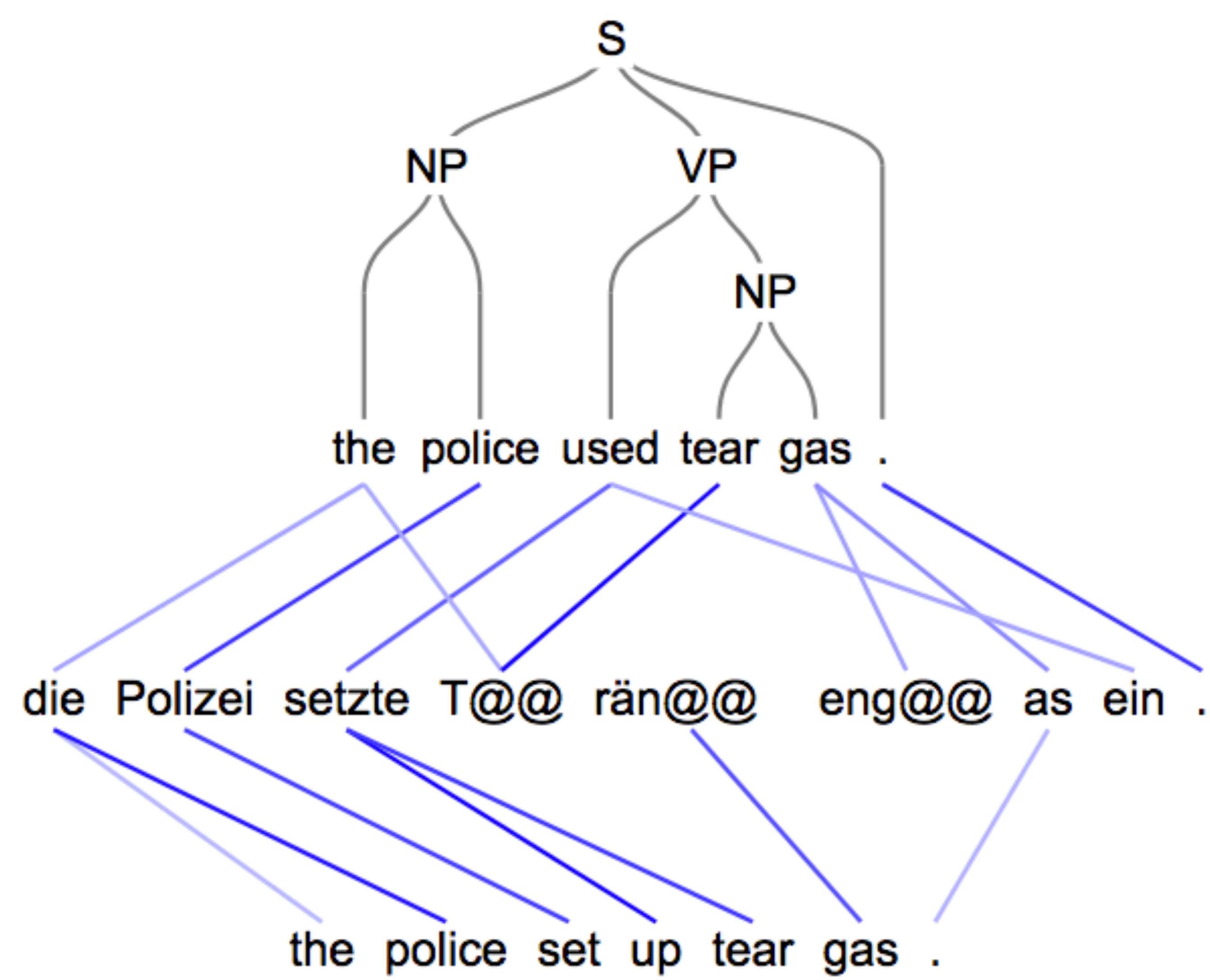
- The syntax-aware model produced more sensible alignments



Where Syntax Helps? Alignments

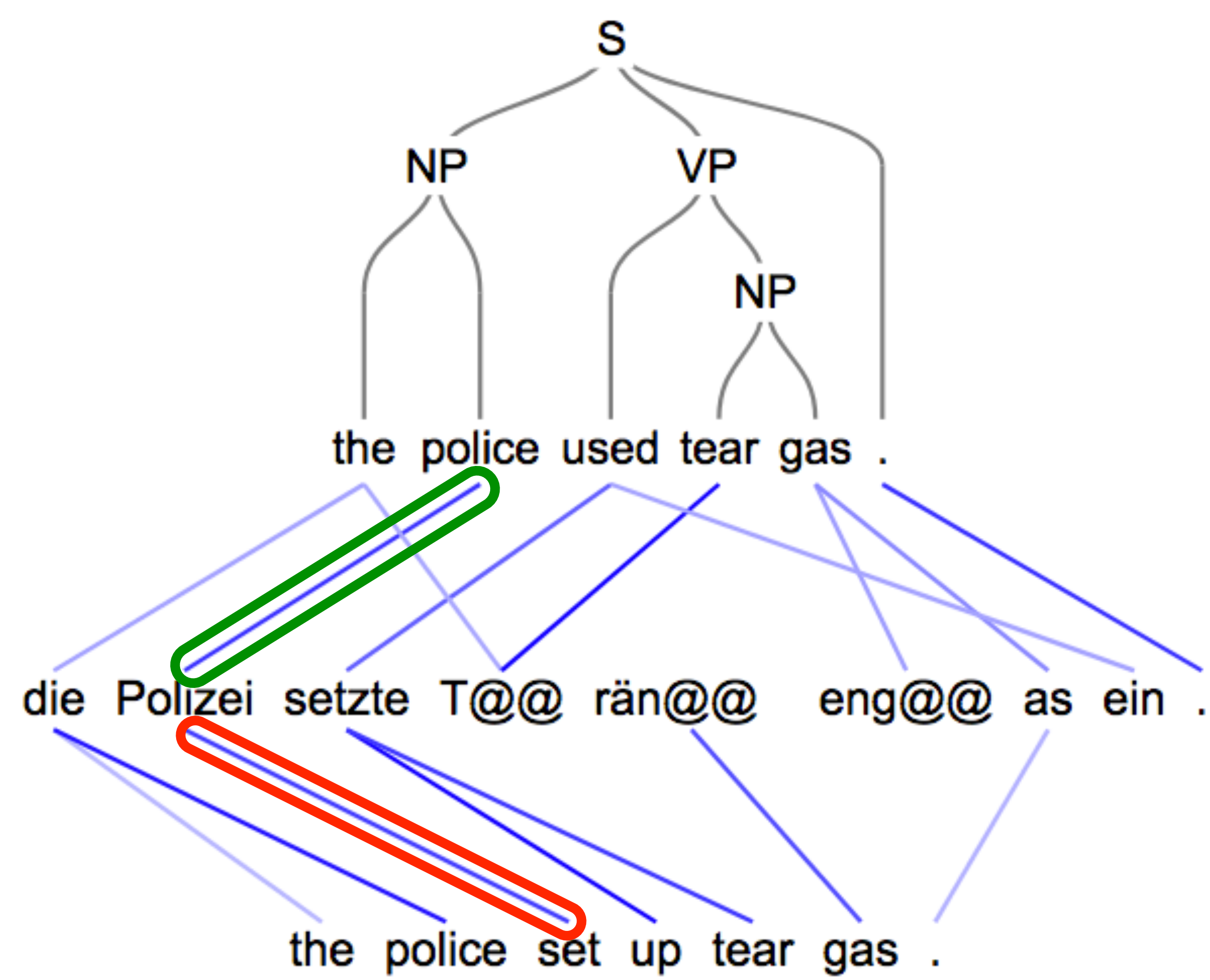
- The attention based model induces soft **alignments** between the source and the target

- The syntax-aware model produces more sensible alignments



Where Syntax Helps? Alignments

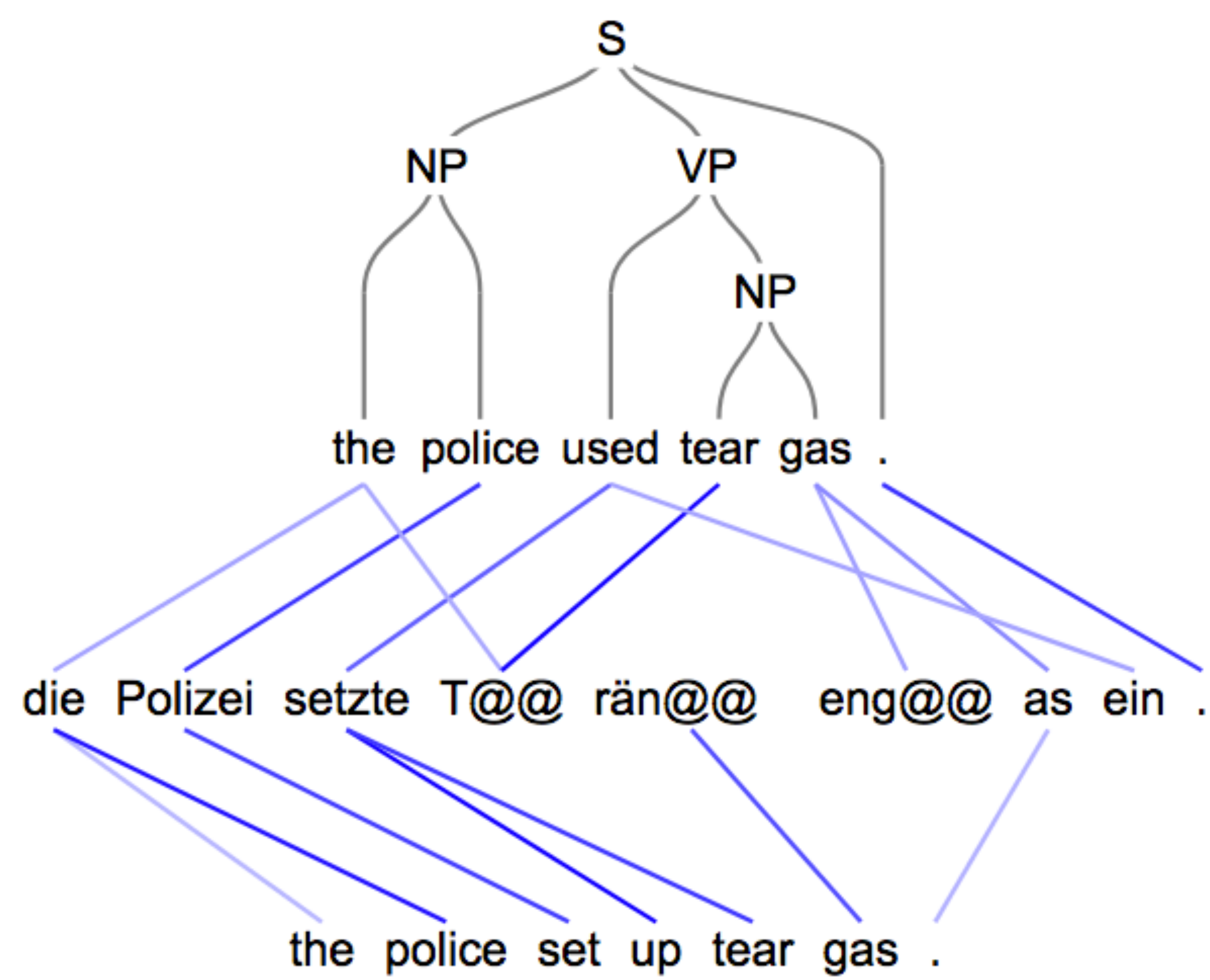
- The attention based model induces soft **alignments** between the source and the target
- The syntax-aware model produces more sensible alignments



Where Syntax Helps? Alignments

- The attention based model induces soft **alignments** between the source and the target

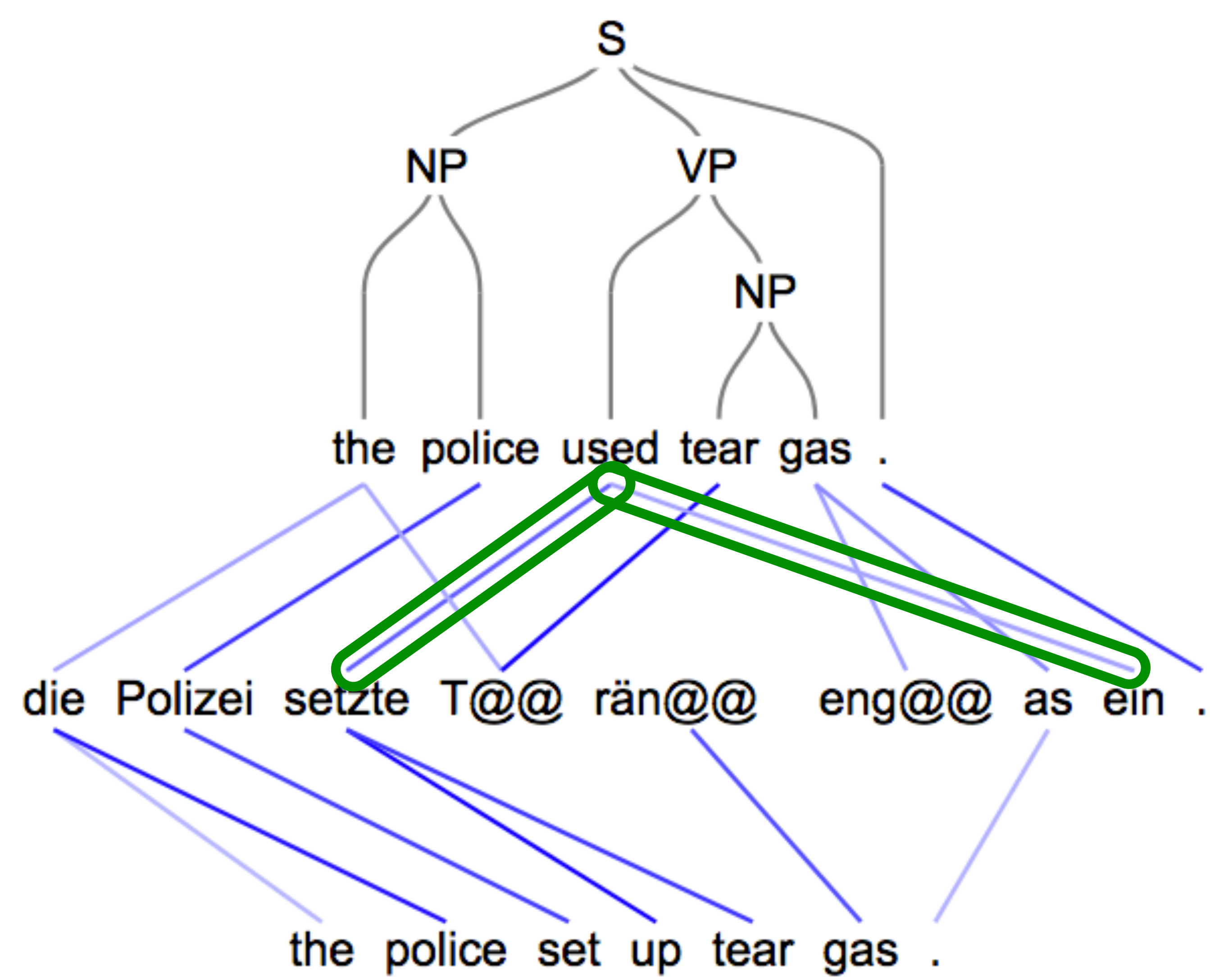
- The syntax-aware model produces more sensible alignments



Where Syntax Helps? Alignments

- The attention based model induces soft **alignments** between the source and the target

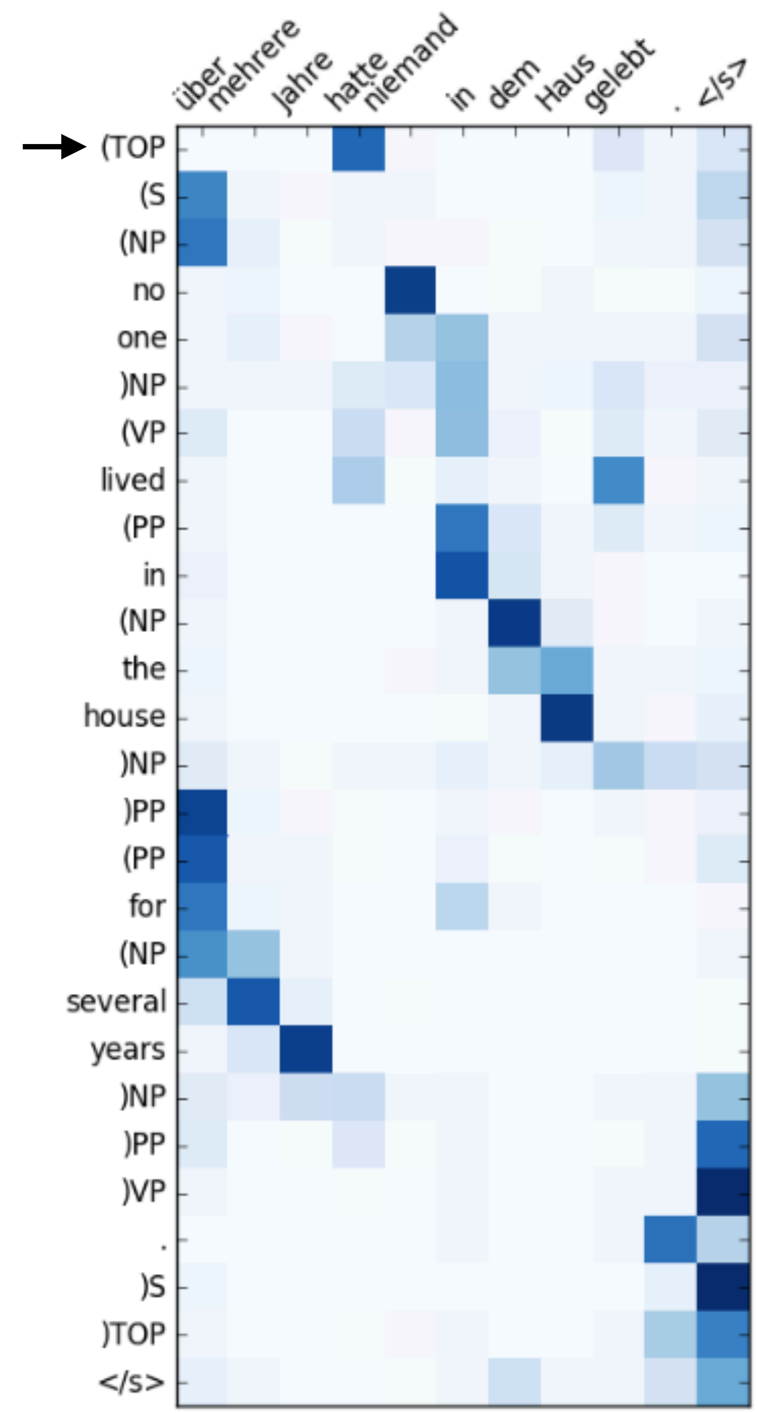
- The syntax-aware model produces more sensible alignments



Attending to Source Syntax

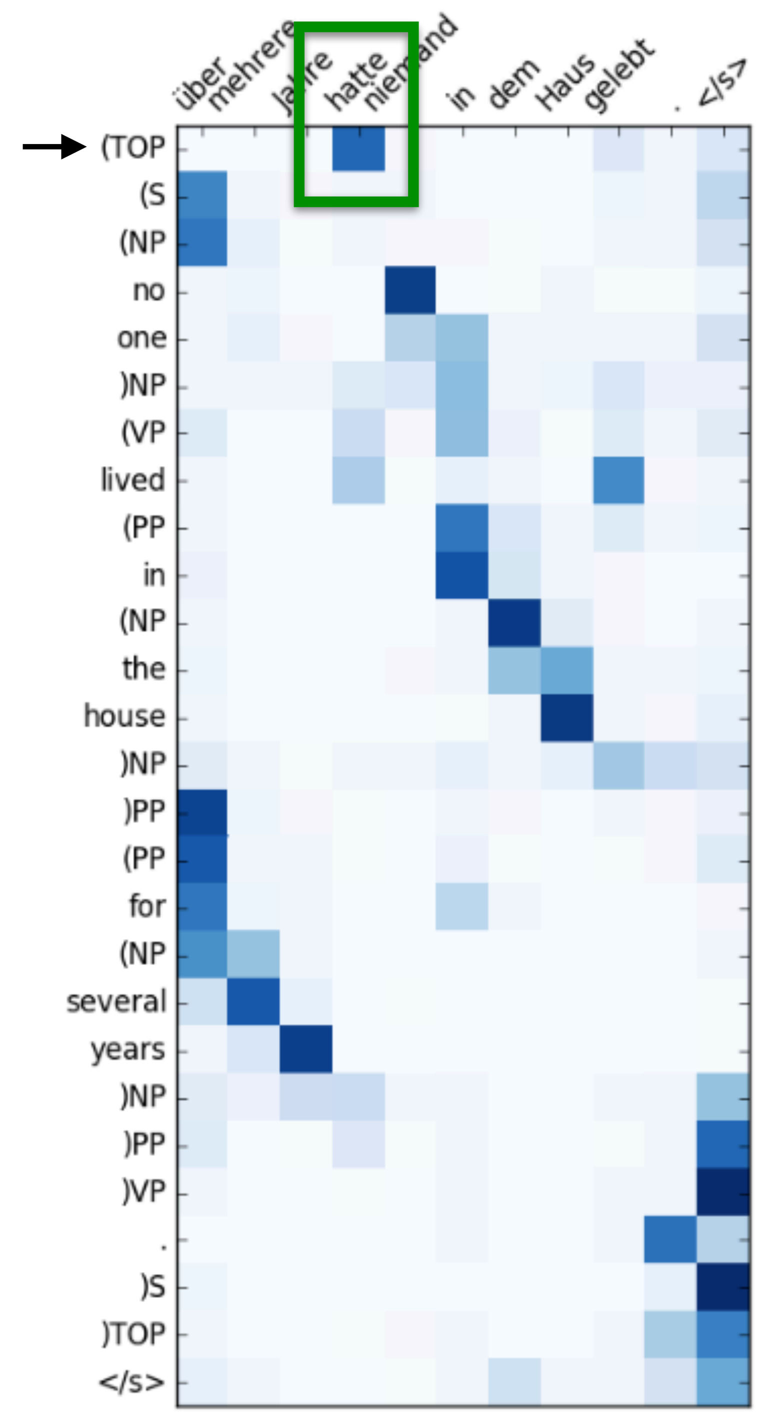
Attending to Source Syntax

- We inspected the attention weights during the production of the tree's **opening brackets**



Attending to Source Syntax

- We inspected the attention weights during the production of the tree's **opening brackets**
- The model consistently attends to the **main verb** ("hatte") or to structural markers (question marks, hyphens…) in the source sentence

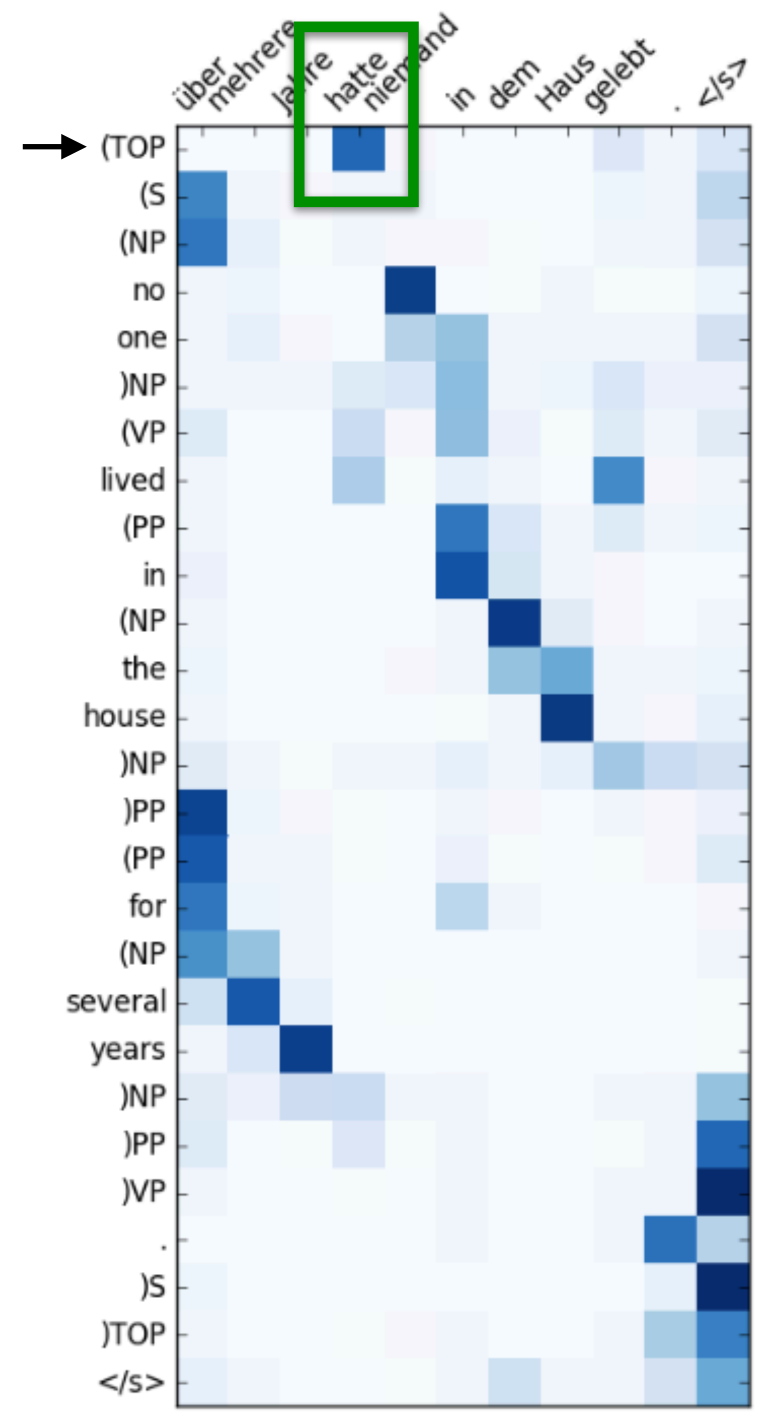


Attending to Source Syntax

- We inspected the attention weights during the production of the tree's **opening brackets**

- The model consistently attends to the **main verb** ("hatte") or to structural markers (question marks, hyphens…) in the source sentence

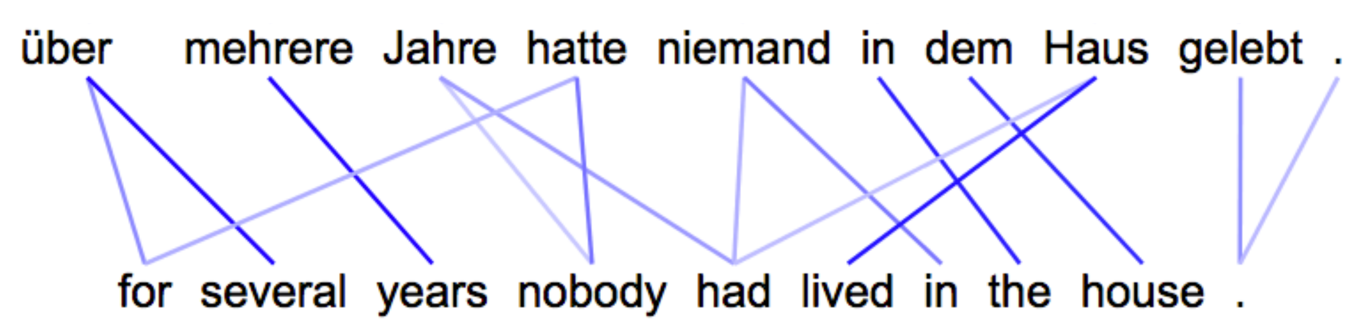
- Indicates the system implicitly learns **source syntax** to some extent (Shi, Padhi and Knight, 2016) and possibly **plans** the decoding accordingly



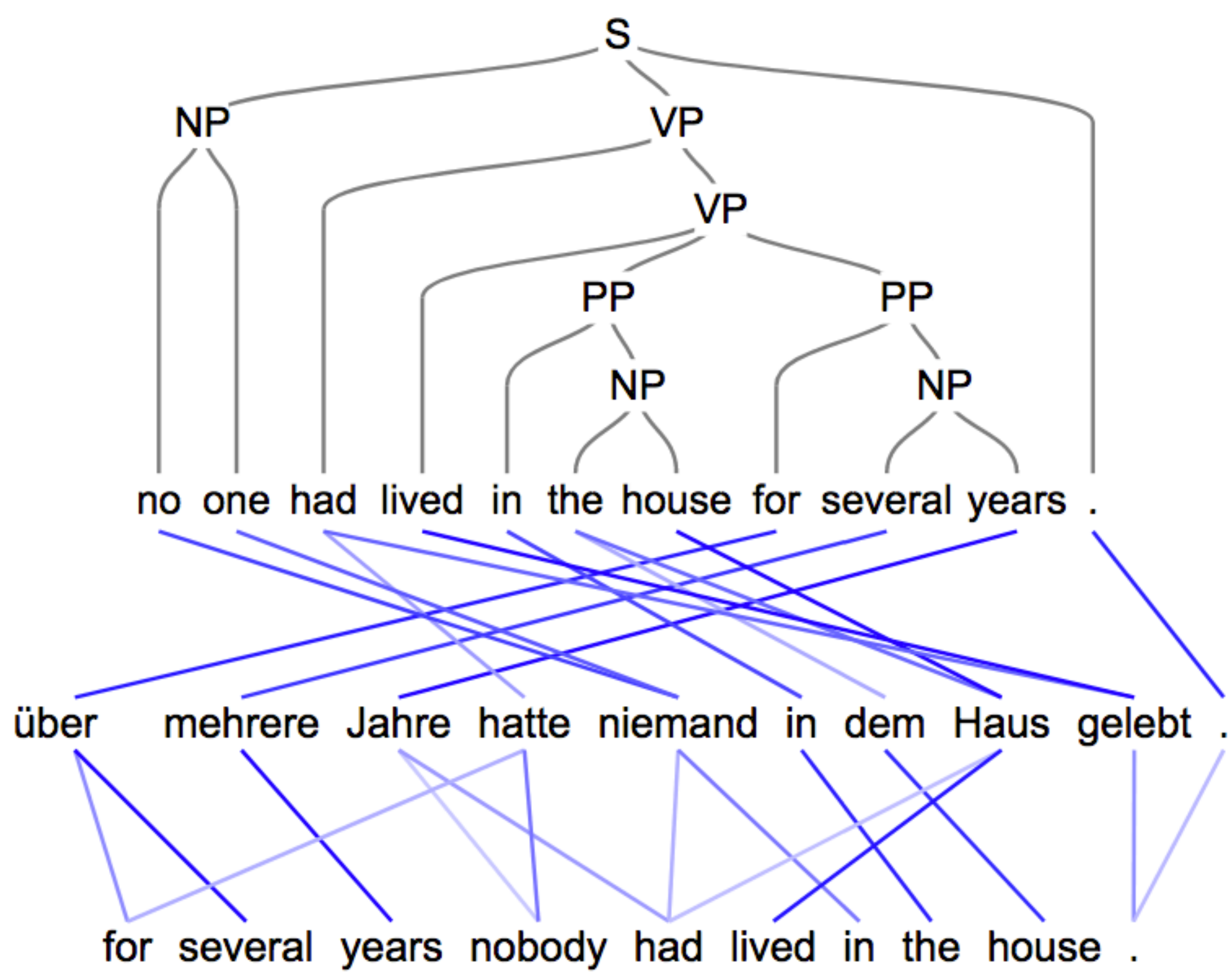
Where Syntax Helps? Structure

über mehrere Jahre hatte niemand in dem Haus gelebt .

Where Syntax Helps? Structure



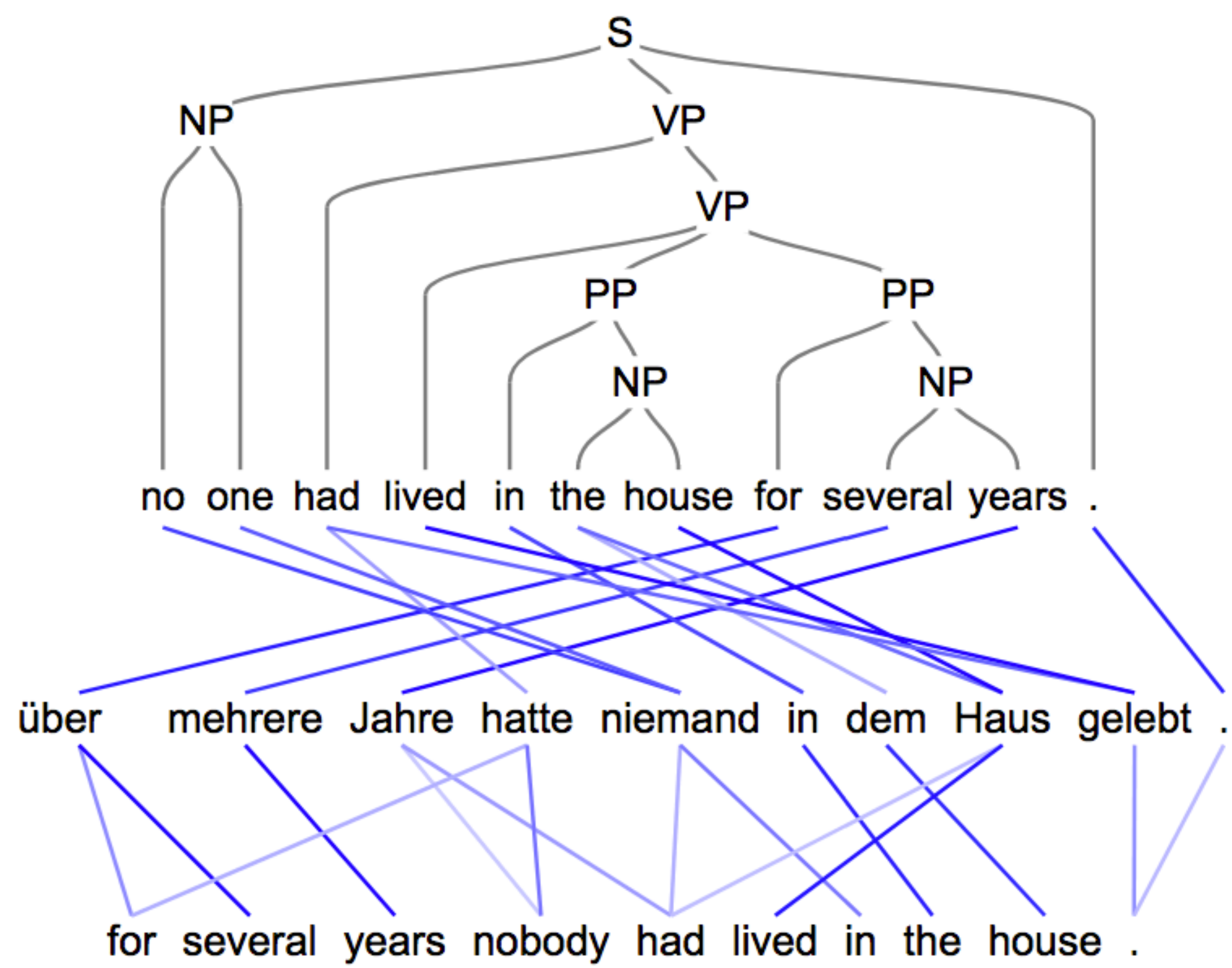
Where Syntax Helps? Structure



Structure (I) - Reordering

Structure (I) - Reordering

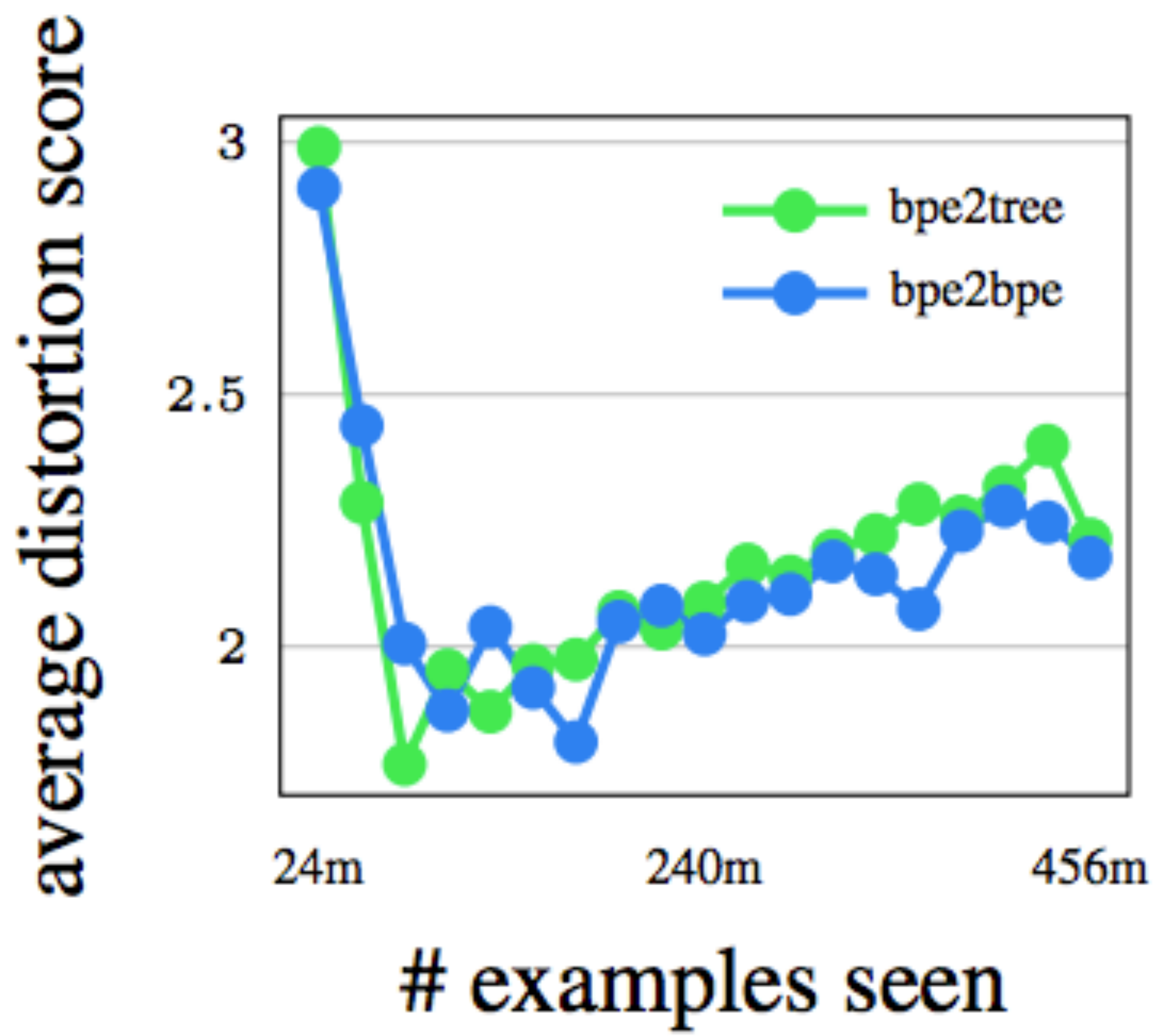
- German to English translation requires a significant amount of **reordering** during translation



Structure (I) - Reordering

- German to English translation requires a significant amount of **reordering** during translation

- Quantifying reordering shows that the syntax-aware system performs **more reordering** during the training process



Structure (I) - Reordering

Structure (I) - Reordering

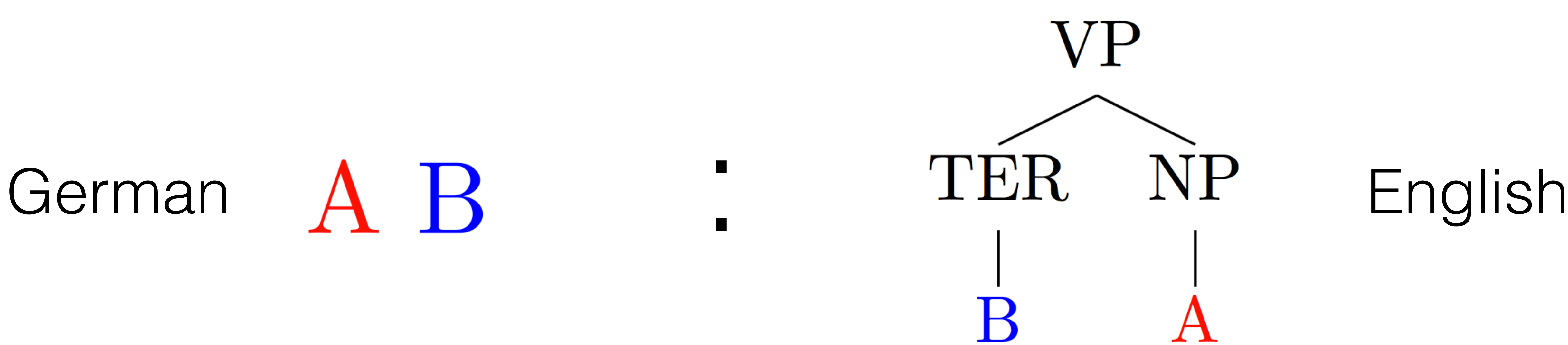
- We would like to **interpret** the increased reordering from a syntactic perspective

Structure (I) - Reordering

- We would like to **interpret** the increased reordering from a syntactic perspective
- We extract **GHKM rules** (Galley et al., 2004) from the dev set using the predicted trees and attention-induced alignments

Structure (I) - Reordering

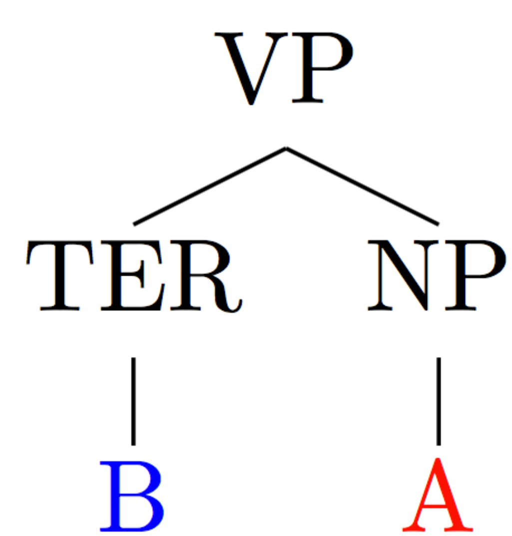
- We would like to **interpret** the increased reordering from a syntactic perspective
- We extract **GHKM rules** (Galley et al., 2004) from the dev set using the predicted trees and attention-induced alignments
- The **most common rules** reveal linguistically sensible transformations, like **moving the verb** from the end of a German constituent to the beginning of the matching English one



Structure (I) - Reordering

- We would like to **interpret** the increased reordering from a syntactic perspective
- We extract **GHKM rules** (Galley et al., 2004) from the dev set using the predicted trees and attention-induced alignments
- The **most common rules** reveal linguistically sensible transformations, like **moving the verb** from the end of a German constituent to the beginning of the matching English one
- More examples in the paper

German A B :



English

Structure (II) - Relative Constructions

Structure (II) - Relative Constructions

- A common linguistic structure is **relative constructions**, i.e. “The XXX **which** YYY”, “A XXX **whose** YYY”…

Structure (II) - Relative Constructions

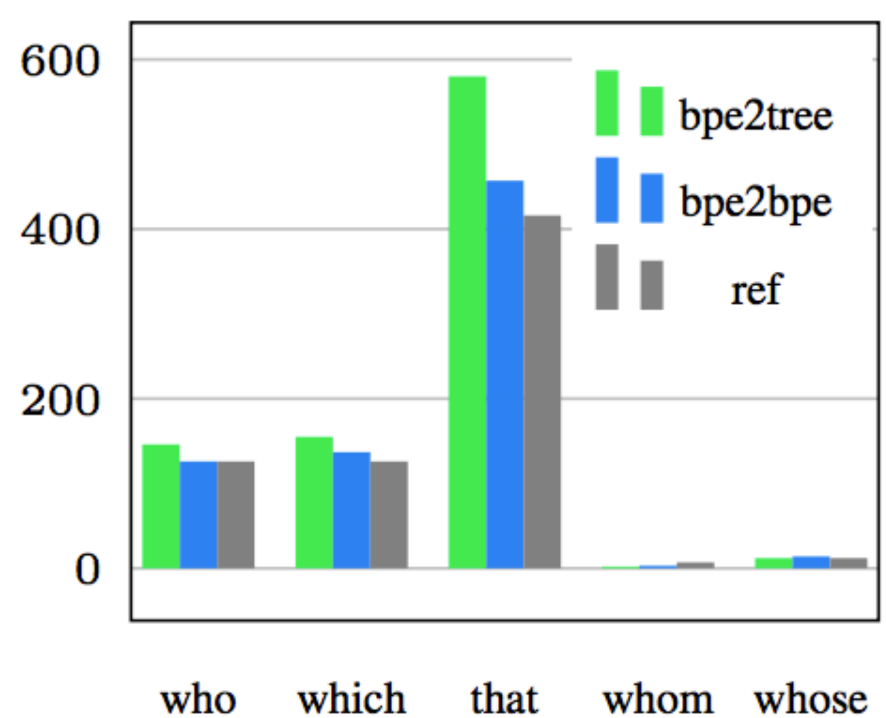
- A common linguistic structure is **relative constructions**, i.e. “The XXX **which** YYY”, “A XXX **whose** YYY”…
- The words that connect the clauses in such constructions are called **relative pronouns**, i.e. “who”, “which”, “whom”…

Structure (II) - Relative Constructions

- A common linguistic structure is **relative constructions**, i.e. “The XXX **which** YYY”, “A XXX **whose** YYY”…

- The words that connect the clauses in such constructions are called **relative pronouns**, i.e. “who”, “which”, “whom”…

- The syntax-aware system produced more relative pronouns due to the syntactic context



Structure (II) - Relative Constructions

Structure (II) - Relative Constructions

Source:

“Guangzhou, das in Deutschland auch Kanton genannt wird…”

Structure (II) - Relative Constructions

Source:

“Guangzhou, das in Deutschland auch Kanton genannt wird…”

Reference:

“Guangzhou, which is also known as Canton in Germany…”

Structure (II) - Relative Constructions

Source:

“Guangzhou, das in Deutschland auch Kanton genannt wird…”

Reference:

“Guangzhou, which is also known as Canton in Germany…”

Syntax-Agnostic:

*“Guangzhou, **also known in Germany**, is one of…*

Structure (II) - Relative Constructions

Source:

“Guangzhou, das in Deutschland auch Kanton genannt wird…”

Reference:

“Guangzhou, which is also known as Canton in Germany…”

Syntax-Agnostic:

*“Guangzhou, **also known in Germany**, is one of…*

Syntax-Based:

*“Guangzhou, **which is also known as the canton in Germany**,…”*

Structure (II) - Relative Constructions

Source:

“Zugleich droht der stark von internationalen Firmen abhängigen Region ein Imageschaden…”

Structure (II) - Relative Constructions

Source:

"Zugleich droht der stark von internationalen Firmen abhängigen Region ein Imageschaden…"

Reference:

"At the same time, the image of the region, which is heavily reliant on international companies…"

Structure (II) - Relative Constructions

Source:

“Zugleich droht der stark von internationalen Firmen abhängigen Region ein Imageschaden…”

Reference:

“At the same time, the image of the region, which is heavily reliant on international companies…”

Syntax-Agnostic:

“At the same time, the ***region's heavily dependent region****…”*

Structure (II) - Relative Constructions

Source:

“Zugleich droht der stark von internationalen Firmen abhängigen Region ein Imageschaden…”

Reference:

“At the same time, the image of the region, which is heavily reliant on international companies…”

Syntax-Agnostic:

*“At the same time, the **region's heavily dependent region**…”*

Syntax-Based:

*“At the same time, the region, **which is heavily dependent on international firms**…”*

Human Evaluation

Human Evaluation

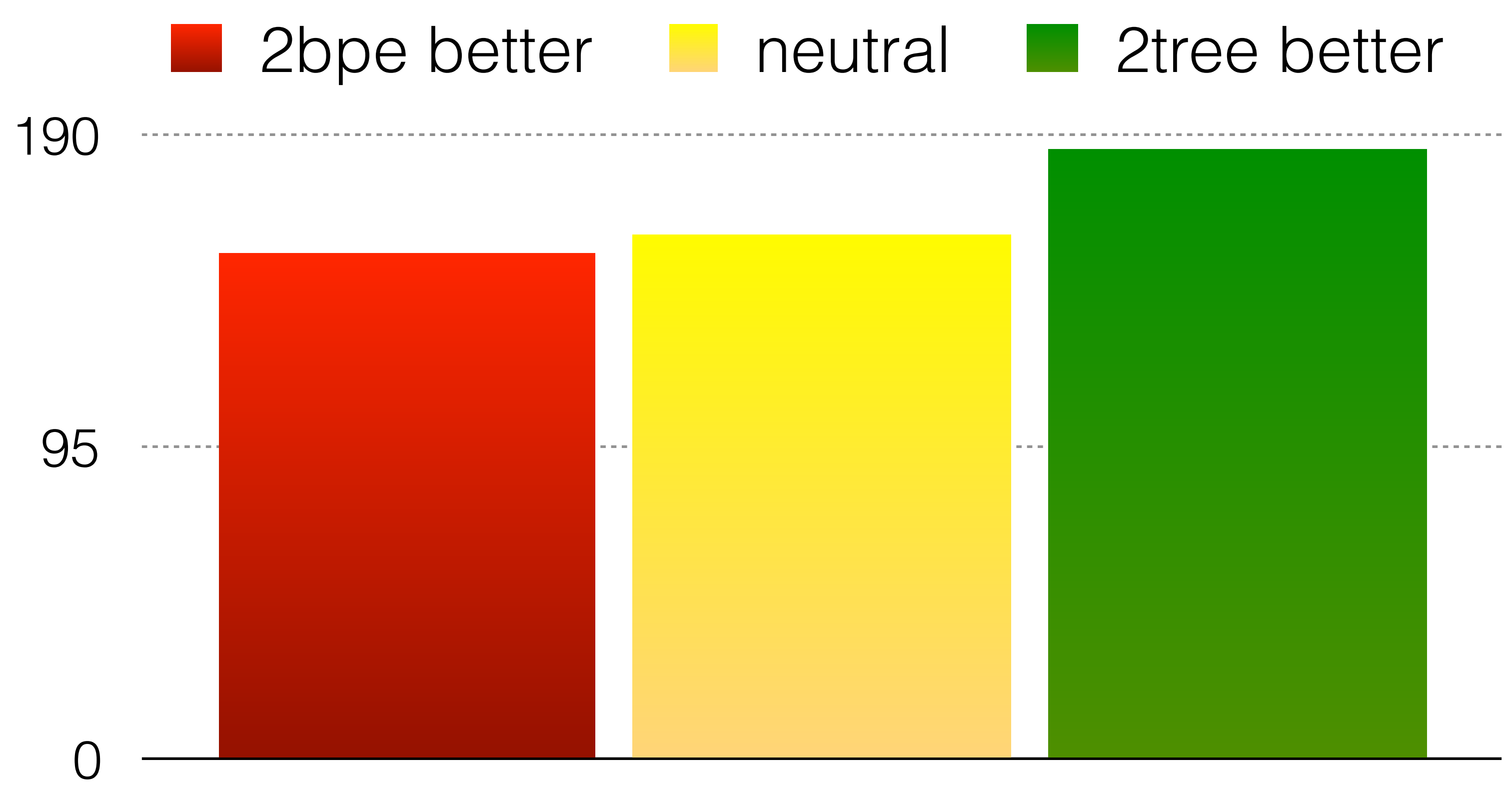
- We performed a small-scale human-evaluation using mechanical turk on the first 500 sentences in newstest 2015

Human Evaluation

- We performed a small-scale human-evaluation using mechanical turk on the first 500 sentences in newstest 2015
- Two turkers per sentence

Human Evaluation

- We performed a small-scale human-evaluation using mechanical turk on the first 500 sentences in newstest 2015
- Two turkers per sentence
- The syntax-aware translations had an advantage over the baseline



Conclusions

Conclusions

- Neural machine translation can clearly **benefit** from target-side syntax

Conclusions

- Neural machine translation can clearly **benefit** from target-side syntax

  Other recent work include:

  - Eriguchi et al., 2017, Wu et al., 2017 (Dependency)

Conclusions

- Neural machine translation can clearly **benefit** from target-side syntax

  Other recent work include:

  - Eriguchi et al., 2017, Wu et al., 2017 (Dependency)
  - Nadejde et al., 2017 (CCG)

Conclusions

- Neural machine translation can clearly **benefit** from target-side syntax

  Other recent work include:

  - Eriguchi et al., 2017, Wu et al., 2017 (Dependency)
  - Nadejde et al., 2017 (CCG)

- A general approach - can be **easily incorporated** into other neural language generation tasks like summarization, image caption generation…

Conclusions

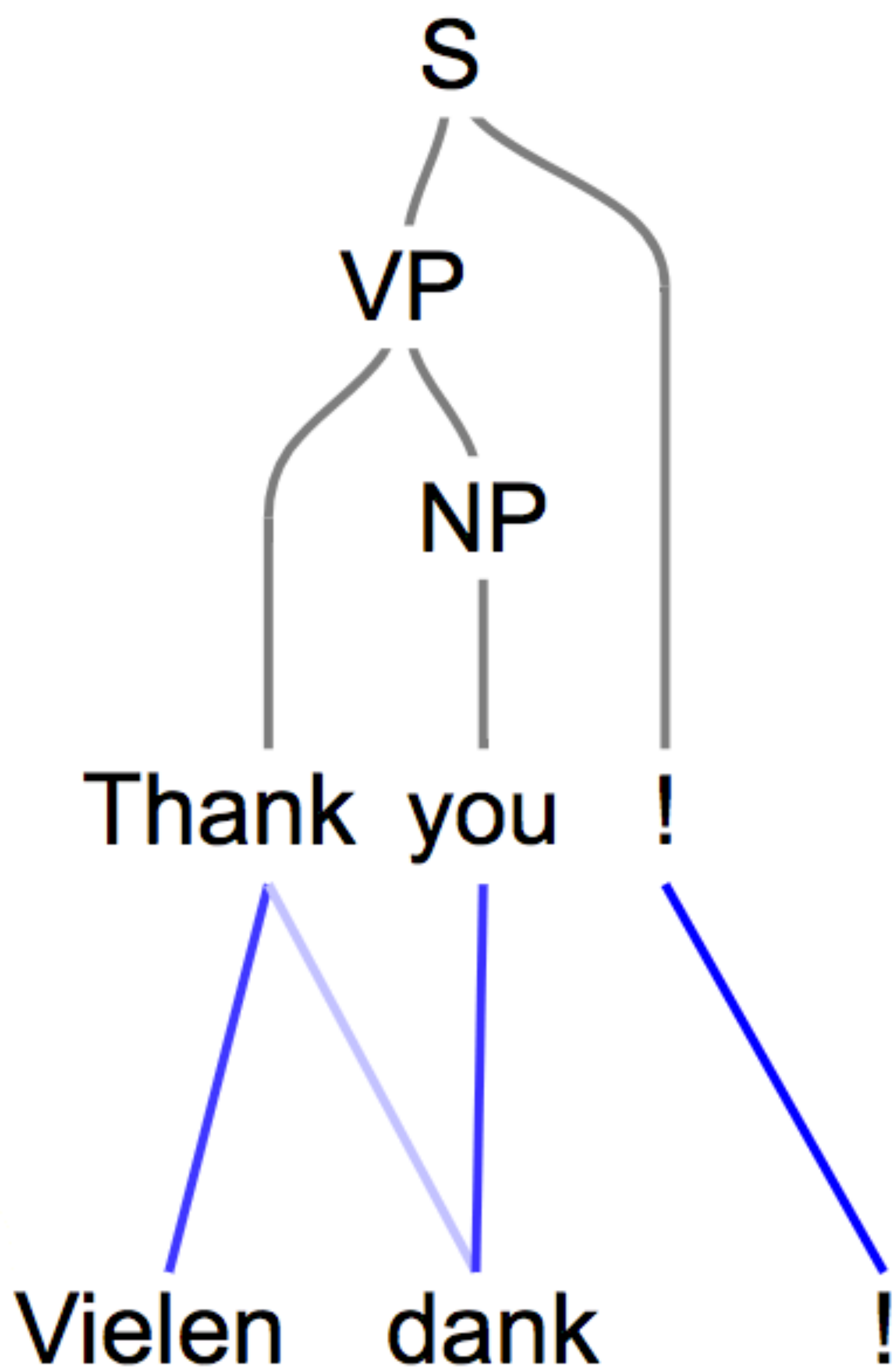
- Neural machine translation can clearly **benefit** from target-side syntax

  Other recent work include:

  - Eriguchi et al., 2017, Wu et al., 2017 (Dependency)
  - Nadejde et al., 2017 (CCG)

- A general approach - can be **easily incorporated** into other neural language generation tasks like summarization, image caption generation…

- Larger picture: **don't throw away your linguistics!** Neural systems can also leverage symbolic linguistic information


S
VP
NP
Thank
you
!
Vielen
dank
!