Dependency-based Sentence Simplification for Increasing Deep LFG Parsing Coverage

Özlem Çetinoğlu, Sina Zarrieß and Jonas Kuhn

IMS, University of Stuttgart

18 July 2013

A German Sentence

From the TIGER Treebank

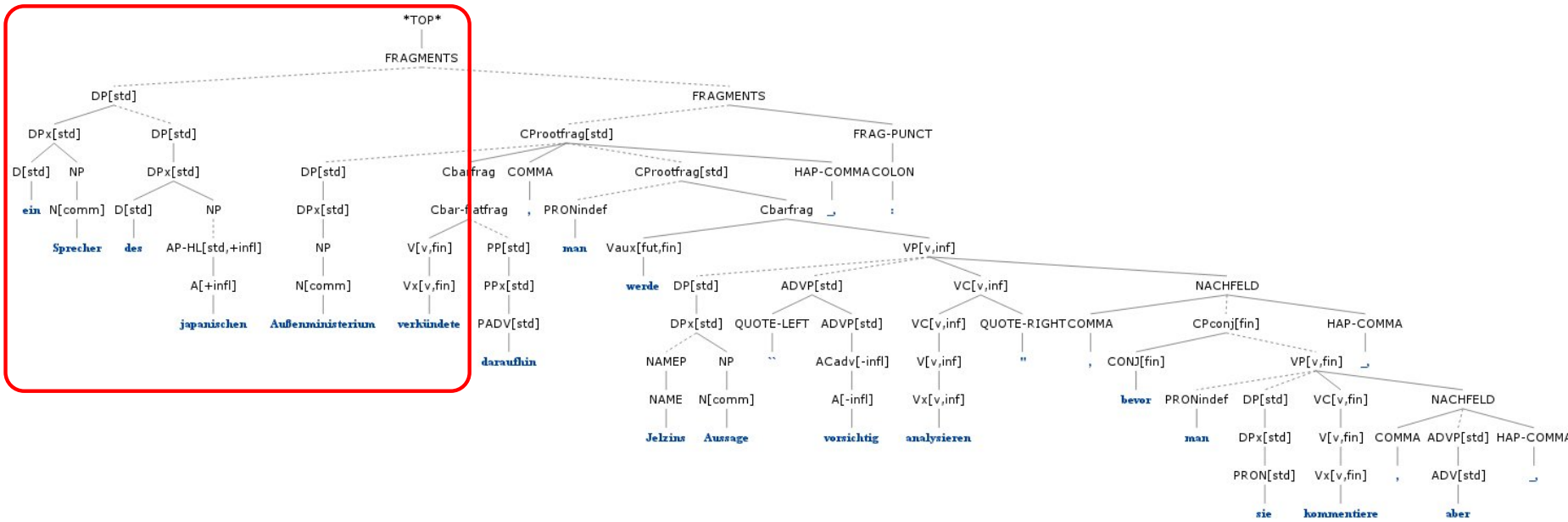
Ein	*Sprecher*	***des***	***japanischen***	***Außenministerium***	*verkündete*	*daraufhin*	,	*man*	*werde*
A	speaker	of the	Japanese	foreign ministry	proclaimed	then	,	one	would

Jelzins	*Aussage*	“	*vorsichtig*	*analysieren*	”	,	*bevor*	*man*	*sie*	*kommentiere*	,	*aber*	:
Yeltsin’s	statement	“	carefully	analyze	”	,	before	one	it	comment	,	but	:

‘A speaker (of the Japanese foreign ministry) then proclaimed that Yeltsin’s statement would be “ carefully analyzed ” , before commenting on it , but :’

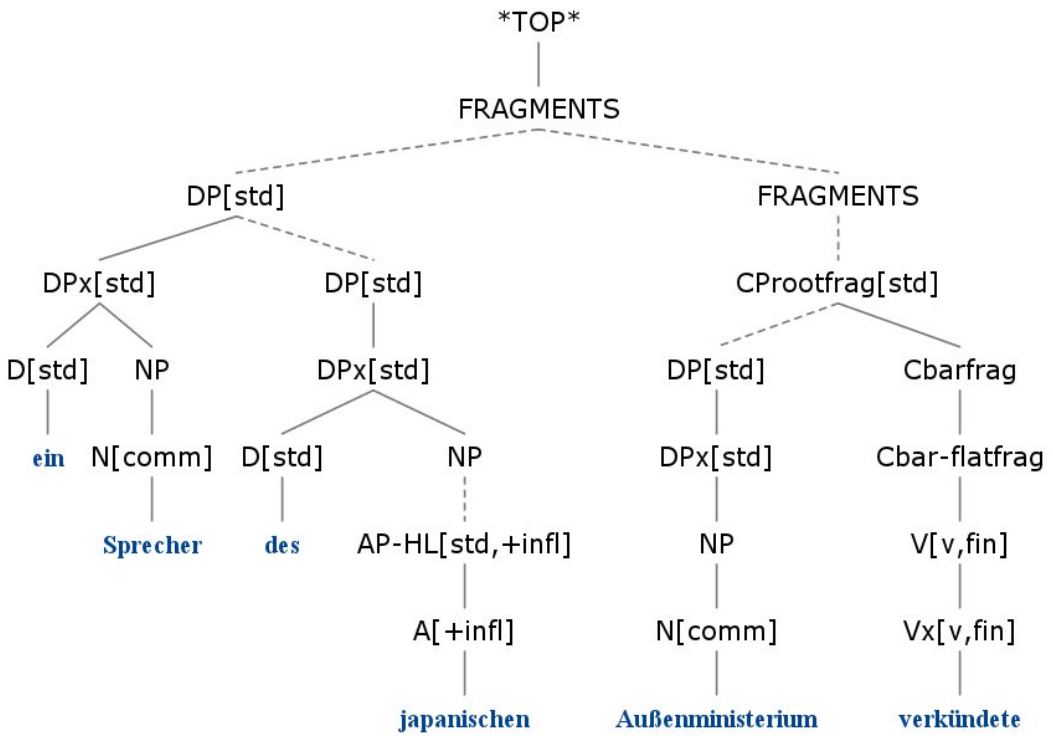
C-structure

Processed with the broad-coverage German LFG grammar

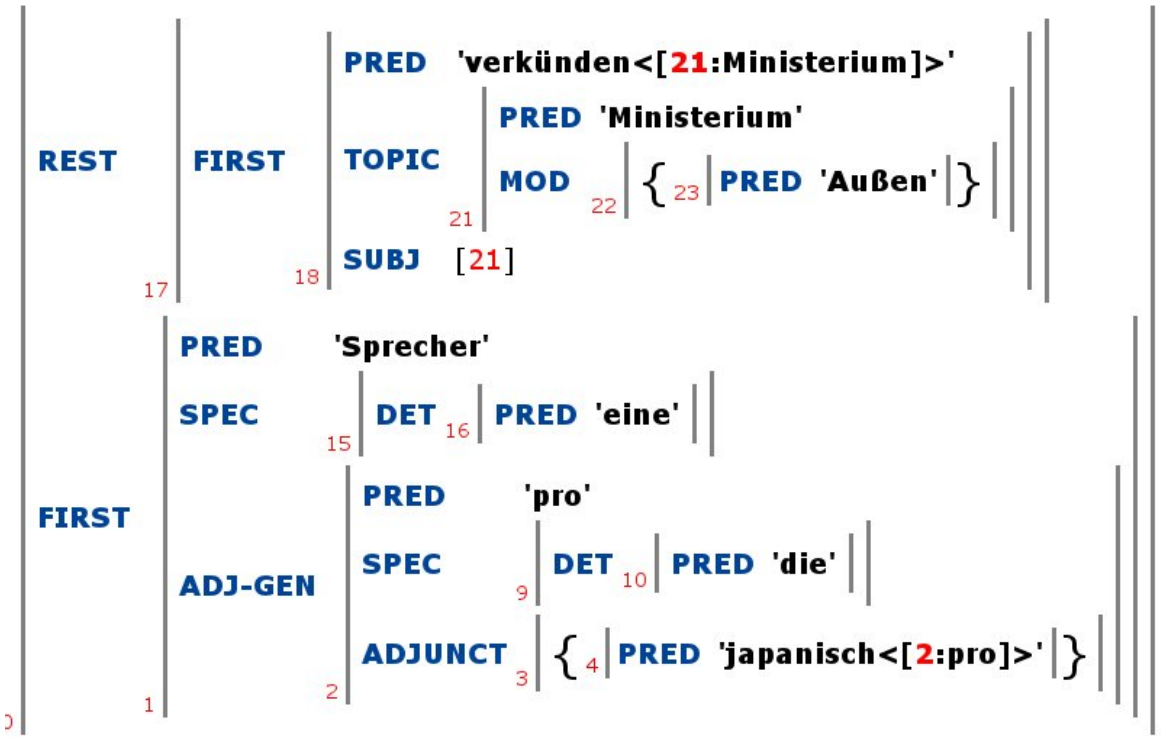


C-structure

Ein	Sprecher	des	japanischen	Außenministerium	verkündete
A	speaker	of the	Japanese	foreign ministry	proclaimed



F-structure



- 'Sprecher' (speaker) does not get any grammatical role
- 'Ministerium' (ministry) is incorrectly analysed as the subject of 'verkünden' (proclaim)

Hand-crafted Grammars

- Wide-coverage
- Deep
- Linguistically motivated

But...

- Cannot reach 100% coverage on unrestricted text
  - Lexical items, idiosyncrasies, rare constructions
  - Ungrammatical material, spelling errors, ...

Hand-crafted Grammars

- We are interested in full parses
  - As gold training data
    - e.g., in XLE parse disambiguation (Forst 2007) and generation ranking (Cahill et al. 2007, Zarrieß et al. 2011)
  - As deep syntactic analyses of raw text
- How can we gain the failed sentences?
  - Can we locate and solve the problem?

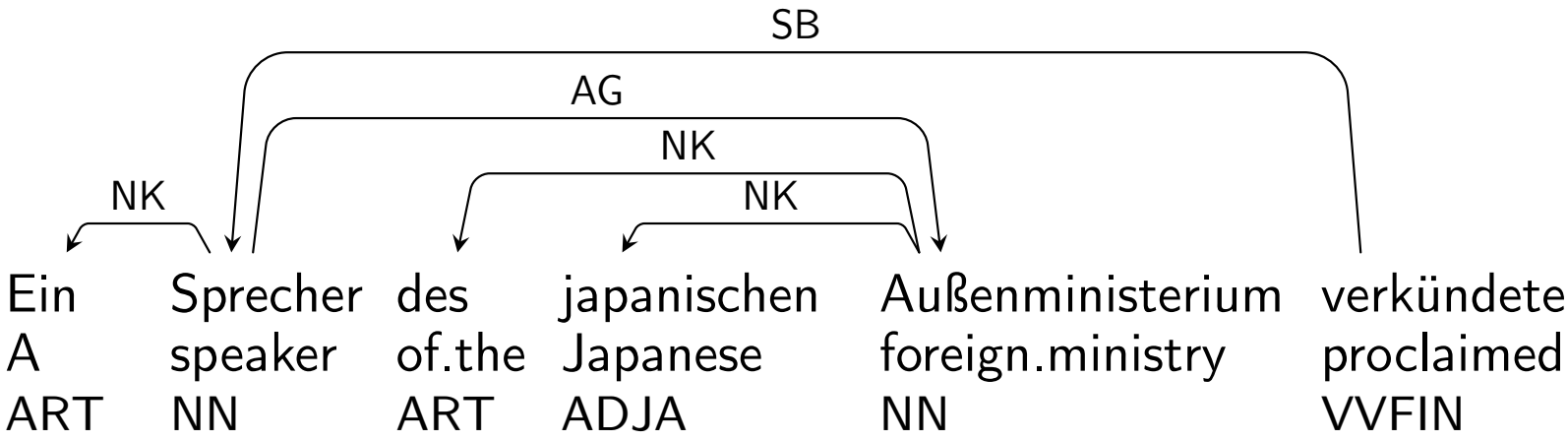
Locating the Problematic Parts

- The problem
  - The genitive marker ‘s’ is missing in ‘Außenministerium’
- After correcting ‘Außenministerium’ to ‘Außenministeriums’
  - Fully connected c-structure
  - Correct arguments in f-structure
- No possible automatic solution
  - Alternative approach?

Locating the Problematic Parts

- More interested in the full parse of core argument structures
- If the problem is located in a modifier phrase
  - Remove it
  - Try to process the sentence again
  - If we get a full parse, core arguments are preserved
- How can we identify modifiers?

- Ein Sprecher **des japanischen Außenministerium** verkündete daraufhin , man werde Jelzins Aussage “ vorsichtig analysieren ” , bevor man sie kommentiere , aber :



SB: Subject AG: Genitive adjunct NK: Noun kernel

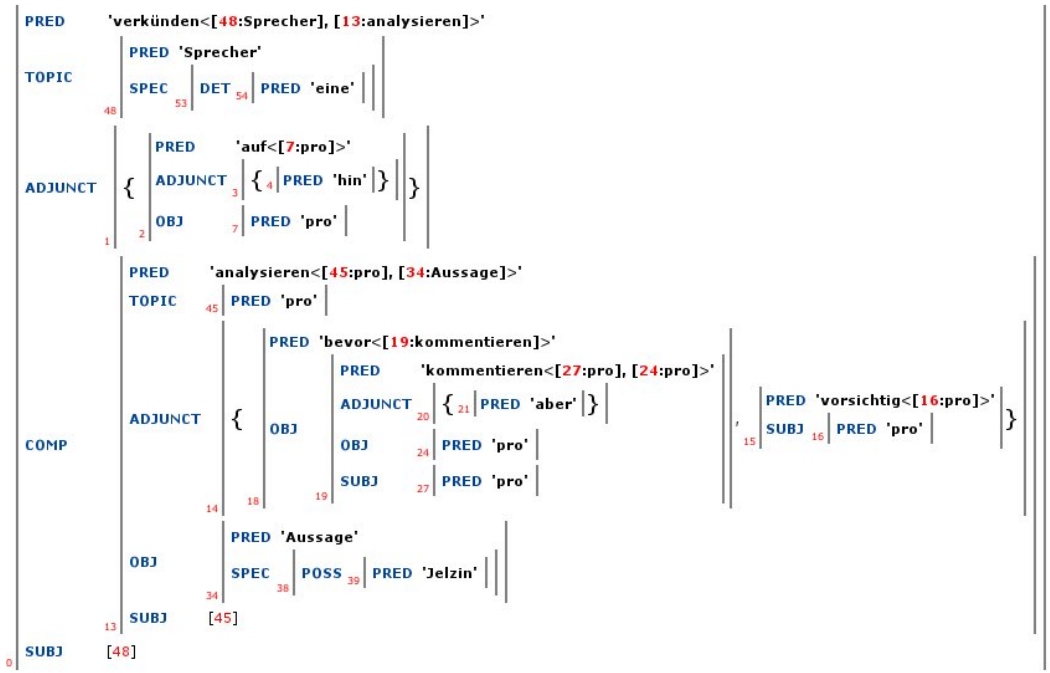
Simplification

- How can we simplify sentences automatically?
- We can utilise their dependency representation
  - Easy/fast to train and to parse with
  - Robust
  - Less sensitive to input errors

Dependency Based Sentence Simplification

- Get the dependency trees of failed sentences
- Delete a subtree from a dependency tree
  - Non-core parts (e.g., appositions, relative clauses)
- Reprocess them

F-structure after Deleting ‘des japanischen Außenministerium’



- ‘Sprecher’ (speaker) is analysed as the subject of ‘verkünden’ (proclaim)
- The complement clause with the head ‘analysieren’ (analyse) is correctly identified

Experiments

- TIGER Treebank (Brants et al. 2002)
- German ParGram Grammar (Rohrer and Forst 2006)
- Sentences 8000 - 10000 are left out as test and development sets
- The remaining sentences are used as the training set

System	sent.	full	failed
TIGER Training	48471	39098	9373
		80.66%	19.34%

- Two sets of experiments
  - Gold dependeny trees
  - Predicted dependency trees

Dependency Based Sentence Simplification

- Convert TIGER gold trees to dependency trees (Seeker and Kuhn, 2012)
- Delete one subtree at a time based on a list of deletable dependencies

Dependency Based Sentence Simplification

AG	genitive adjuncts
APP	appositions
JU	discourse marker-like
MNR	PP adjuncts (in noun phrases)
MO	modifiers
NG	negation
PAR	head of parenthesis
PG	possessive PP adjuncts
PH	placeholders (e.g. German Vorfeld es)
PNC	proper noun components
RC	relative clauses
RE	infinite clauses attached to nominals
SBP	PP subjects in passive
UC	inside foreign language phrases
VO	vocatives
NK	noun kernels (only when they are adjuncts, or subordinate conjunctions with sentence)
DA	datives (could be free)

Dependency Based Sentence Simplification

- Convert TIGER gold trees to dependency trees (Seeker and Kuhn, 2012)
- Delete one subtree at a time based on a list of deletable dependencies
- Apply a set of punctuation correction rules
- The number of candidates depends on the number deletable dependencies of a sentence
- In total, there are 52867 candidates (5.6 candidates per sentence)
- Process all shorter candidates with XLE

Dependency Based Sentence Simplification

- Further simplification: Instead of deleting one subtree, delete all possible subtree combinations
- On average there are 924 candidates per sentence
- For sentences with more than 10 candidates, take the shortest 10 as candidates
- Remove punctuation from the shortest candidate and add it as the 11th candidate
- The average number of candidates per sentence drops to 8.1

Dependency Based Sentence Simplification

Original:

- Ein Sprecher [$_{AG}$ des japanischen Außenministerium] verkündete [$_{MO}$ daraufhin] , man werde Jelzins Aussage “ [$_{MO}$ vorsichtig] analysieren ” , [$_{MO}$ bevor man sie kommentiere ,] [$_{MO}$ aber] :

A speaker [of the Japanese foreign ministry] [then] proclaimed that Yeltsin's statement would be “ [carefully] analyzed ” , [before commenting on it ,] [but] :

Simplified:

- Ein Sprecher [$_{AG}$] verkündete [$_{MO}$], man werde Jelzins Aussage “ [$_{MO}$] analysieren ” [$_{MO}$] [$_{MO}$]:
- Ein Sprecher [$_{AG}$] verkündete daraufhin , man werde Jelzins Aussage “ [$_{MO}$] analysieren ” [$_{MO}$] [$_{MO}$]:
- Ein Sprecher [$_{AG}$] verkündete [$_{MO}$], man werde Jelzins Aussage “ vorsichtig analysieren ” [$_{MO}$] [$_{MO}$]:

Baseline Approach

- N-gram based sentence simplification
  - Uses the parsability metric of van Noord (2004)
- Parsability of a word:

$$P(w) = \frac{C(w|OK)}{C(w)}$$

- Parsability of a word sequence:

$$P(w_i \dots w_j) = \frac{C(w_i \dots w_j|OK)}{C(w_i \dots w_j)}$$

N-gram Based Sentence Simplification

- Get n-grams of failed sentences
- Calculate their number of occurrence
  - in failed sentences
  - in the whole treebank
- And calculate the parsability of n-grams in failed sentences
- Delete zero parsability n-grams
- Reprocess them

Note that this approach does not ensure the grammaticality of a simplified sentence or the preservation of argument structure

N-gram Based Sentence Simplification

Original:

- Ein Sprecher des japanischen Außenministerium verkündete daraufhin , man werde Jelzins Aussage “ vorsichtig analysieren ” , bevor man sie kommentiere , aber :

A speaker of the Japanese foreign ministry then proclaimed that Yeltsin's statement would be “ carefully analyzed ” , before commenting on it , but :

Simplified:

- ~~Ein~~ Sprecher des japanischen Außenministerium verkündete daraufhin , man werde Jelzins Aussage “ vorsichtig analysieren ” , bevor man sie ~~kommentiere~~ , aber :
- ~~Ein~~ Sprecher des japanischen Außenministerium verkündete daraufhin , man werde Jelzins Aussage “ vorsichtig ~~analysieren ”~~ , bevor man sie kommentiere , aber :
- ~~Ein~~ Sprecher des ~~japanischen Außenministerium verkündete daraufhin ,~~ man ~~werde Jelzins Aussage “ vorsichtig analysieren ” , bevor man sie kommentiere ,~~ aber :

N-gram Based Sentence Simplification

- For each failed sentence, calculate n-gram parsability scores (n=1,2,3)
- Delete n-grams with zero-parsability
- If there are no n-grams with zero parsability, delete the n-gram with the lowest parsability
- Apply a set of punctuation correction rules
- In total, there are 26822 candidates
- Process all shorter candidates with XLE

XLE Coverage

Sentences with at least one full parse

System	sent.	full parses
TIGER Training	48471	39098 (80.66%)
n-gram deletion	9373	2893 (30.87%)
1 subtree shorter	9373	3367 (35.92%)
10 shortest	9373	4607 (49.83%)
1 subtree shorter + 10 shortest	9373	4909 (52.37%)

- The upper limit of simplified sentences with a full parse is 8462 (90.28%) because 911 sentences are not simplifiable at all
  - No deletable subtrees

Checking Simplified F-structures

- Accuracy of the simplification system

- Assume
  - Gold: Time flies like an arrow
  - Not parsable
- Remove the modifier
- Reprocess
  - Time flies

Checking Simplified F-structures

- How can we check the accuracy of the simplification system?
  - Check against the TIGER trees
- Technically intricate: different annotation/representation

- We apply Forst's (2007) approach to simplified sentences
- Check if XLE parses are compatible with TIGER trees

TIGER-Compatible F-structures

System	sent.	full parses	TIGER-compatible
TIGER Training	48471	39098	11931 (30.53%)
1 subtree shorter	9373	3367	665 (19.75%)
10 shortest	9373	4607	2345 (50.90%)
1 subtree shorter + 10 shortest	9373	4909	2381 (48.50%)

- percentages: The ratio of TIGER-compatible parses to full parses

Getting Dependencies for Raw Text

- Parse the TIGER sentences with a statistical dependency parser (Bohnet 2010)
- Lemma, POS, and morphological features are also predicted
- All systems are trained on the TIGER data by using cross-validation

Getting Dependencies for Raw Text

System	sent.	full parses
Predicted		
1 subtree shorter	9373	3211 (34.26%)
10 shortest	9373	4346 (46.37%)
1 subtree shorter + 10 shortest	9373	4738 (50.55%)
Gold		
1 subtree shorter	9373	3367 (35.92%)
10 shortest	9373	4607 (49.83%)
1 subtree shorter + 10 shortest	9373	4909 (52.37%)

A Closer Look into Failed Sentences

What cannot be parsed? (among all 9373 failed sentences)

Parsability	Count	n-gram
0.000	11	Befreiungstiger von Tamil
0.000	11	CDU / CSU
0.000	17	# ski #
0.000	29	90 / Die
0.000	31	/ dpa /
0.000	34	afp / dpa
0.000	38	# (...
0.000	40	# (rtr
0.000	41	dpa / rtr
0.000	95	# (dpa

# denotes sentence boundary

A Closer Look into Failed Sentences

What helped when deleted? (in the 1 subtree shorter simplification)

Count	Phrase
189	sich
82	nicht
40	so
34	auch
22	ihm
18	Immer
16	rund
15	nur
15	aber

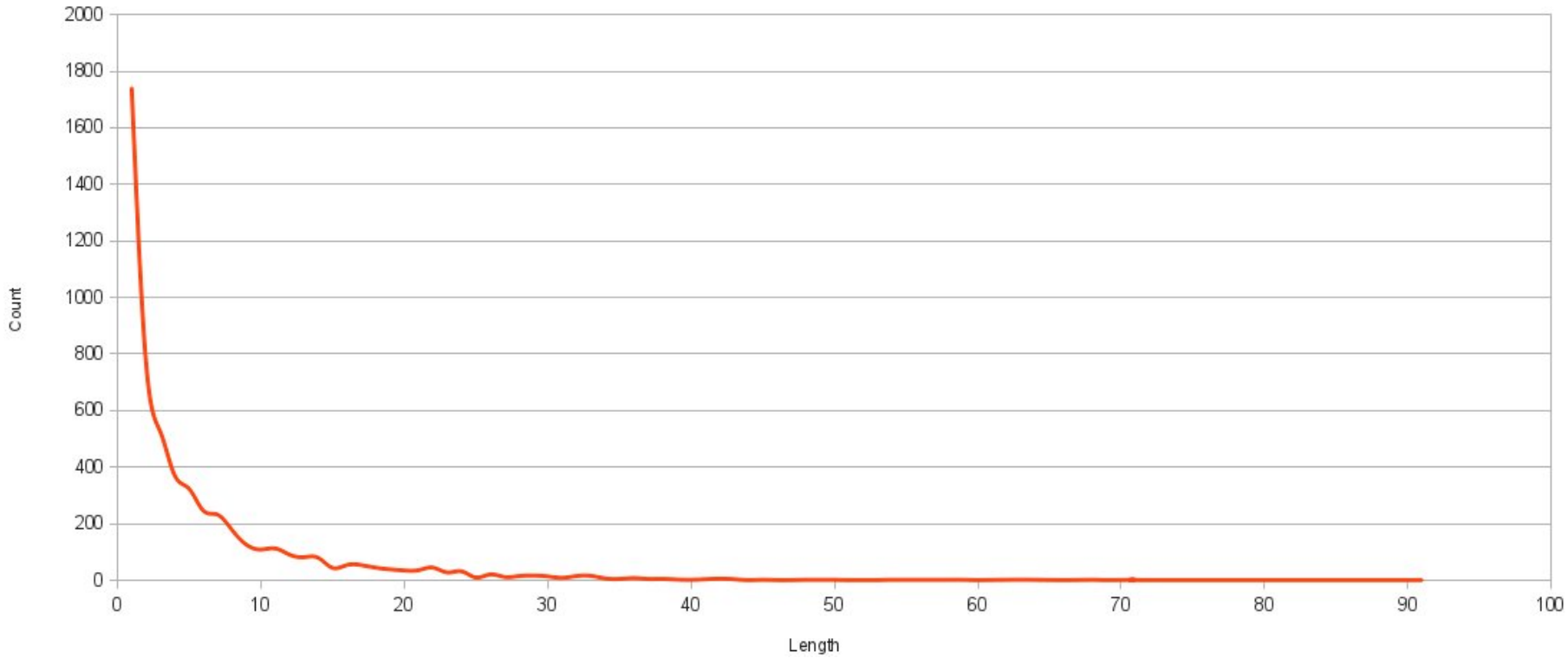
An Example with Deleting ‘nicht’

Daß	*es*	*so*	*einfach*	***nicht***	*ist*	,	*weiß*	*natürlich*	*auch*	*der*	*CDU-Politiker*	.
That	it	so	easy	**not**	is	,	knows	naturally	also	the	CDU politician	.

‘Of course the CDU politician also knows it is **not** so easy.

A Closer Look into Failed Sentences

How much we have to delete? (in the 1 subtree shorter simplification)



1 token: 1742 2 tokens: 770 3 tokens: 525 4 tokens: 371 5+ tokens: 2205

Conclusion

- We present a dependency based simplification approach
  - to improve the full parse coverage
  - while we ensure grammaticality
  - and preserve core parts
- Experiments on the TIGER treebank show
  - we gain 52.37% of the failed sentences with this approach
  - and 48.50% of the gained sentences have the accurate parse
  - when we apply the system to predicted dependencies, results are comparable to the gold setting

Future Work

- An improved simplification approach
  - Dependency subtree deletion based on the parsability of n-grams
- Using the extended set of compatible f-structures
  - XLE parse disambiguation
  - Generation reranking
- Utilising the deep syntactic representations
  - As features of a dependency parser

Thanks!

Questions?