On the relationship between (Enhanced) Universal Dependencies and HPSG

Gosse Bouma

Virtual Leuven/Berlin/Seattle, August, 18, 2020

University of Groningen

Summary

An existential crisis?

Is there a need for formal grammar in the age of BERT?

- Statistical models pre-trained on huge amounts of raw text and fine-tuned on modest amounts of annotated data perform at amazing levels, outperforming (statistical) rule-based approaches
- So where does that leave theories such as HPSG?

Hypothesis 1

- Treebank annotation is shallow and misses many features essential for full interpretation. Richer representations will require more linguistic guidance

Hypothesis 2

- Linguistic theory provides constraints & generalisations that can improve accuracy of neural models

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible

On the linguistic adequacy of Enhanced UD

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible
- Enhanced UD is able to capture most of the information that is present in formal grammar models (HPSG, LFG)

On the linguistic adequacy of Enhanced UD

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible
- Enhanced UD is able to capture most of the information that is present in formal grammar models (HPSG, LFG)
- Slightly different formats are possible as well:
  - Surface-Syntactic UD (Gerdes et al, 2017, 2019): capture (phrasal) distributional generalisations (function words are heads) repartitioning dependency relations, surface vs deep syntax

On the linguistic adequacy of Enhanced UD

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible
- Enhanced UD is able to capture most of the information that is present in formal grammar models (HPSG, LFG)
- Slightly different formats are possible as well:
  - Surface-Syntactic UD (Gerdes et al, 2017, 2019): capture (phrasal) distributional generalisations (function words are heads) repartitioning dependency relations, surface vs deep syntax
- Semantic Interpretation is underdeveloped area:
  - Reddy, 2017, Gotham and Haug, 2018

What about Phrase Strucure?

State of the Art Dependency Parsers

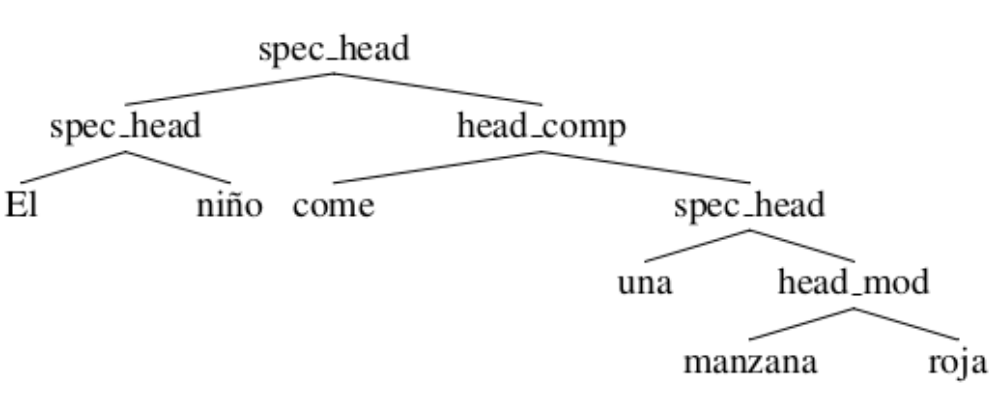
English PTB (Anderson and Gómez-Rodríguez, 2020)

	Parser	UAS	LAS	Sent/sec CPU
graph	HPSG (Zhou and Zhao, 2019)	**96.09**	**94.68**	
graph	BIST-Graph (Kiperwasser and Goldberg, 2016)	93.10	91.00	**80**
graph	Biaffine (Dozat and Manning, 2017)	95.74	94.08	~22
graph	Biaffine (Zhang, 2020)	95.74	94.07	57
transition	Pointer-TD (Ma et al., 2018)	95.87	94.19	10
transition	Pointer-LR (Fernández-González and Gómez-Rodríguez, 2019)	**96.04**	**94.43**	23
transition	UUParser (de Lhoneux et al., 2017)	94.63	92.77	42
transition	BIST-Transition (Kiperwasser and Goldberg, 2016)	93.90	91.90	76
transition	CM (Chen and Manning 2014)	91.80	89.60	**654**

(*Linear Neural Parsing and Hybrid Enhancement for Enhanced Universal Dependencies*, Attardi *et al*, IWPT 2020)

Can we apply this trick more widely?

Spanish Ancora corpus

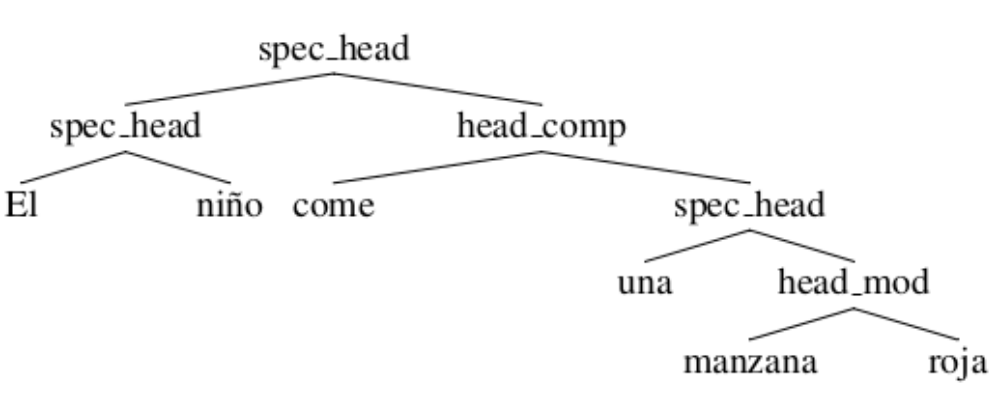


the boy eats a red apple

Chiruzzo and Wonserver, 2020,Statistical Deep Parsing for Spanish

Can we apply this trick more widely?

Spanish Ancora corpus



the boy eats a red apple

Chiruzzo and Wonserver, 2020,Statistical Deep Parsing for Spanish

UD v2.5

- 158 treebanks, 100 treebanks are converted from non-UD data:

Relations | annotated manually in non-UD style, automatically converted to UD

Introduction

Introduction

SOTA in syntactic parsing

Recently, neural parsers without using any grammar rules significantly outperform conventional statistical grammar-based ones for the reason that neural networks, especially recurrent models (e.g, Bi-LSTM), are adept in capturing long range contextual information (anonymous, under review)

SOTA for Dutch Lassy Small UD Treebank

	System	LAS
Early days (2017)	Alpino, van Noord 2007	**84.31**
	ParseySaurus, Alberti et al 2017	80.53
	Parsey's Cousins, Andor et al 2016	78.08
	Easy-first, Kiperwasser 2016	77.16
CONLL 2017 (*NNs catch up*)	Stanford	**87.71**
	IMS (Stuttgart)	86.86
	HIT-SCIR (Harbin)	86.85
	NAIST SATO (Nara)	85.03

An existential crisis?

Is there a need for formal grammar in the age of BERT?

- Statistical models pre-trained on huge amounts of raw text and fine-tuned on modest amounts of annotated data perform at amazing levels, outperforming (statistical) rule-based approaches
- So where does that leave theories such as HPSG?

An existential crisis?

Is there a need for formal grammar in the age of BERT?

- Statistical models pre-trained on huge amounts of raw text and fine-tuned on modest amounts of annotated data perform at amazing levels, outperforming (statistical) rule-based approaches
- So where does that leave theories such as HPSG?

Hypothesis 1

- Treebank annotation is shallow and misses many features essential for full interpretation. Richer representations will require more linguistic guidance

An existential crisis?

Is there a need for formal grammar in the age of BERT?

- Statistical models pre-trained on huge amounts of raw text and fine-tuned on modest amounts of annotated data perform at amazing levels, outperforming (statistical) rule-based approaches
- So where does that leave theories such as HPSG?

Hypothesis 1

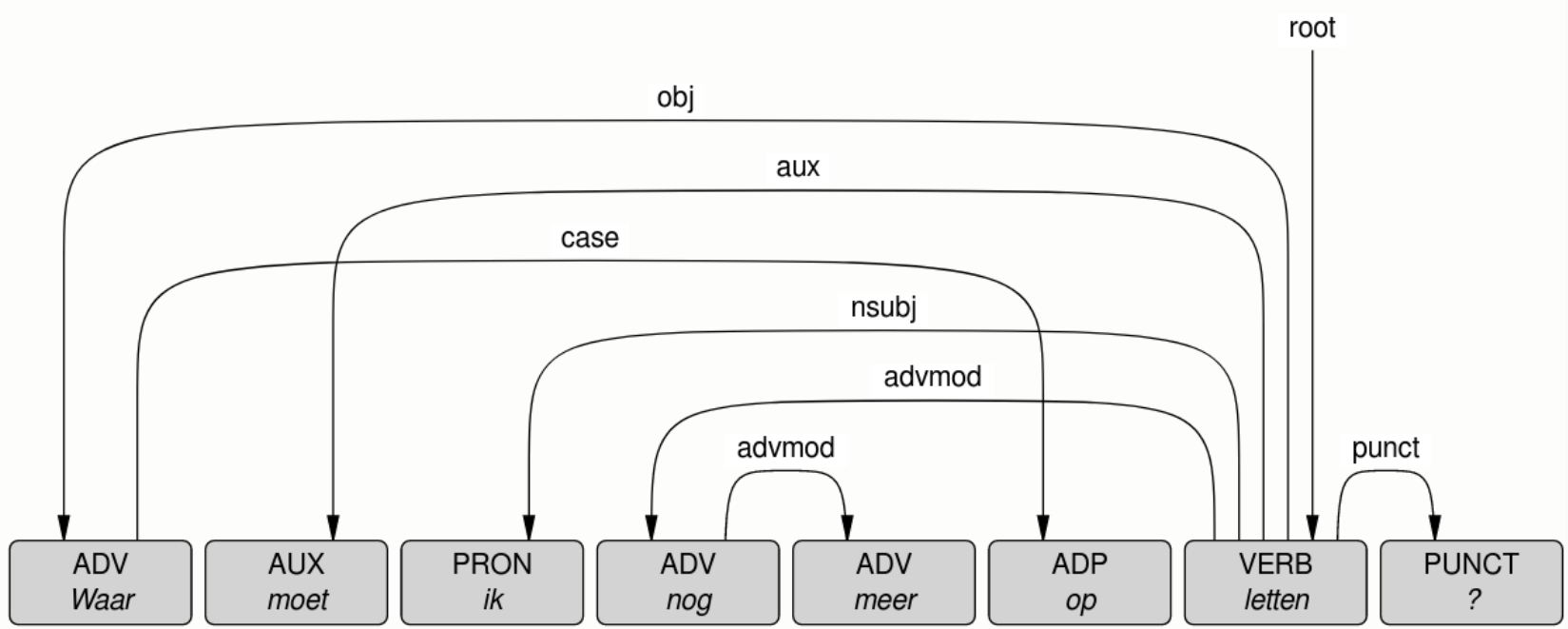
- Treebank annotation is shallow and misses many features essential for full interpretation. Richer representations will require more linguistic guidance

Hypothesis 2

- Linguistic theory provides constraints & generalisations that can improve accuracy of neural models

Enhanced Universal Dependencies

Universal Dependencies in a nutshell



Where must I still more at look ?

What else should I pay attention to?

Universal Dependencies in a nutshell

Dependency Annotation

- Single annotation scheme for all languages
  - uniform inventory of POS, features and depedency labels
- Maximize cross-lingual consistency
  - Primacy of content words over function words

Universal Dependencies in a nutshell

Dependency Annotation

- Single annotation scheme for all languages
  - uniform inventory of POS, features and depedency labels
- Maximize cross-lingual consistency
  - Primacy of content words over function words

Formal properties

- Terminals annotated with lemma, POS, morphosyntactic features
- Labeled head-dependent edges between terminals
- Annotation is an (unordered) tree with single root
  - Non-projective edges for crossing and long distance dependencies
  - No empty nodes

Zeman, Nivre et al, UD treebanks v1.0 - v2.6

Enhanced Universal Dependencies

Motivation

- Universal Dependencies is an easy-to-use annotation standard for many languages
- But formal constraints make it hard to capture some phenomena correctly (i.e. control, coordination, ellipsis)
- and do not fully support downstream applications (e.g. Question Answering, Information Extraction)

Enhanced Universal Dependencies

Motivation

- Universal Dependencies is an easy-to-use annotation standard for many languages
- But formal constraints make it hard to capture some phenomena correctly (i.e. control, coordination, ellipsis)
- and do not fully support downstream applications (e.g. Question Answering, Information Extraction)

Enhanced Universal Dependencies

- Annotation is a (cyclic) graph
- Terminals can be dependents of multiple heads
- 'Empty' nodes are allowed

http://universaldependencies.org/u/overview/enhanced-syntax.html

Enhanced Universal Dependencies

Coordination Dependency relations are propagated for (some) dependents and predicates

Control and raising The external subject of *xcomp* is explicitly marked

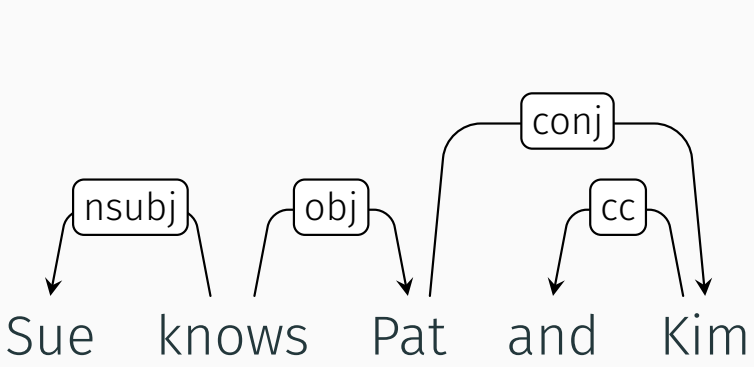
Gapping and Ellipsis Empty tokens in the input represent missing predicates

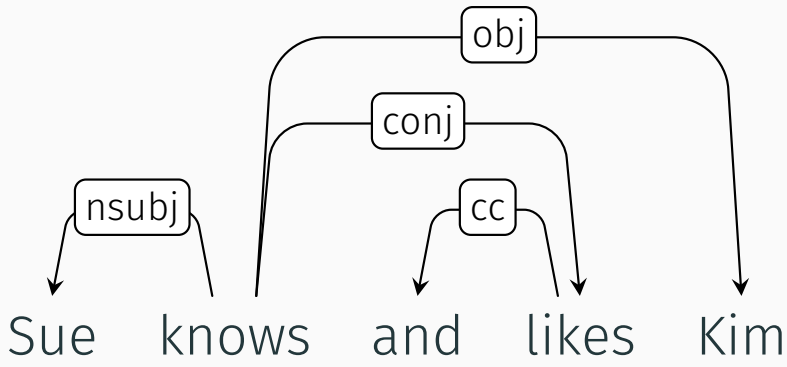
Relative clauses Antecedent noun is a dependent of some node in the relative clause (thus introducing a cycle)

Case information Selected dependents become *rel:case* where *case* is the lemma of a case/marker/cc dependent and/or the case feature of the dependent

Coordination

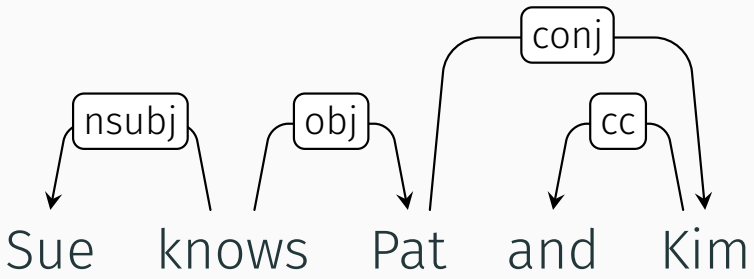
Basic UD

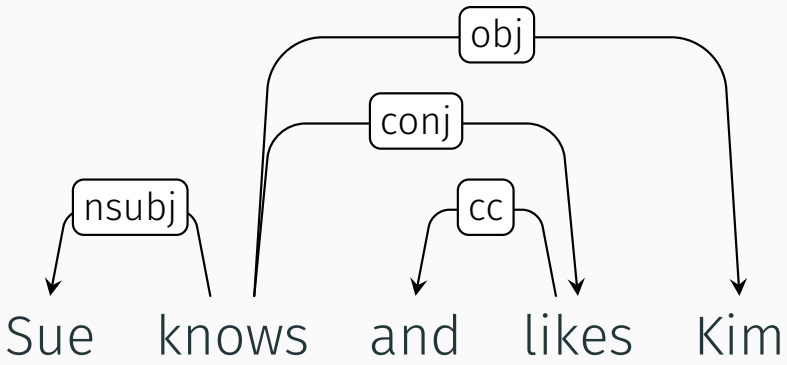




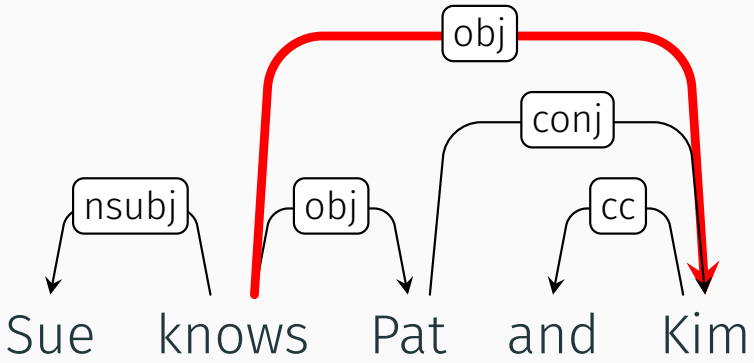
Coordination

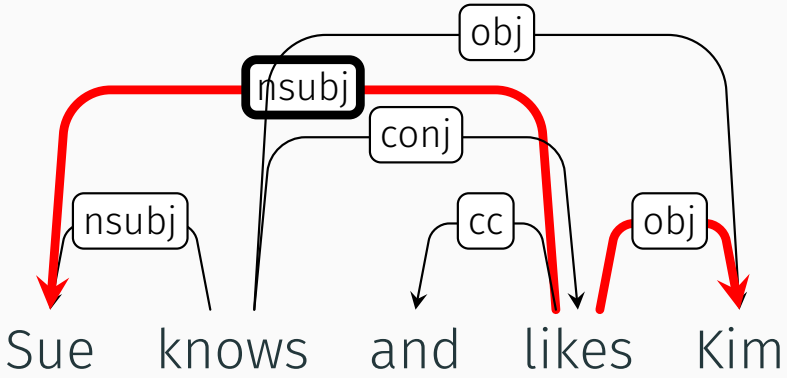
Basic UD





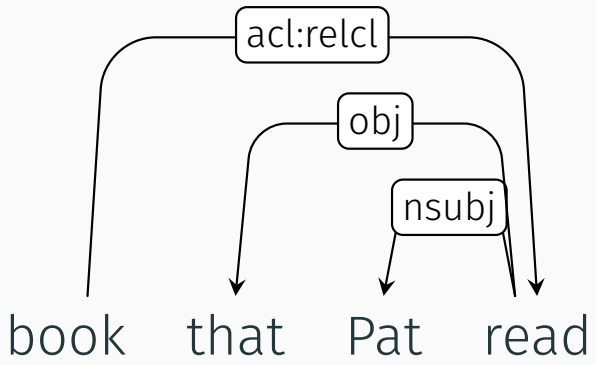
Enhanced UD





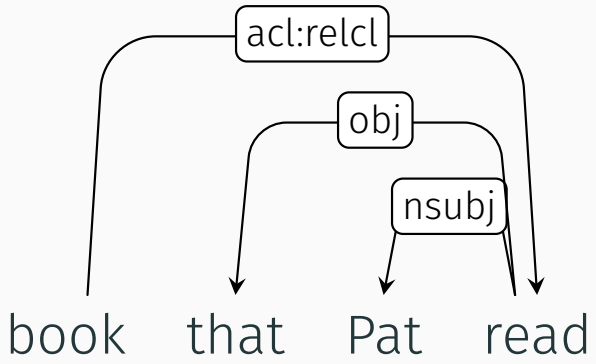
Relatives

Basic UD

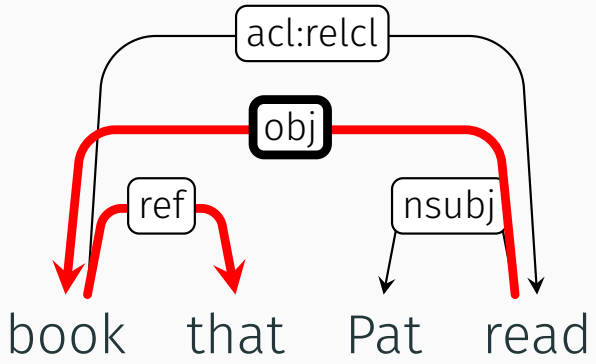


Relatives

Basic UD

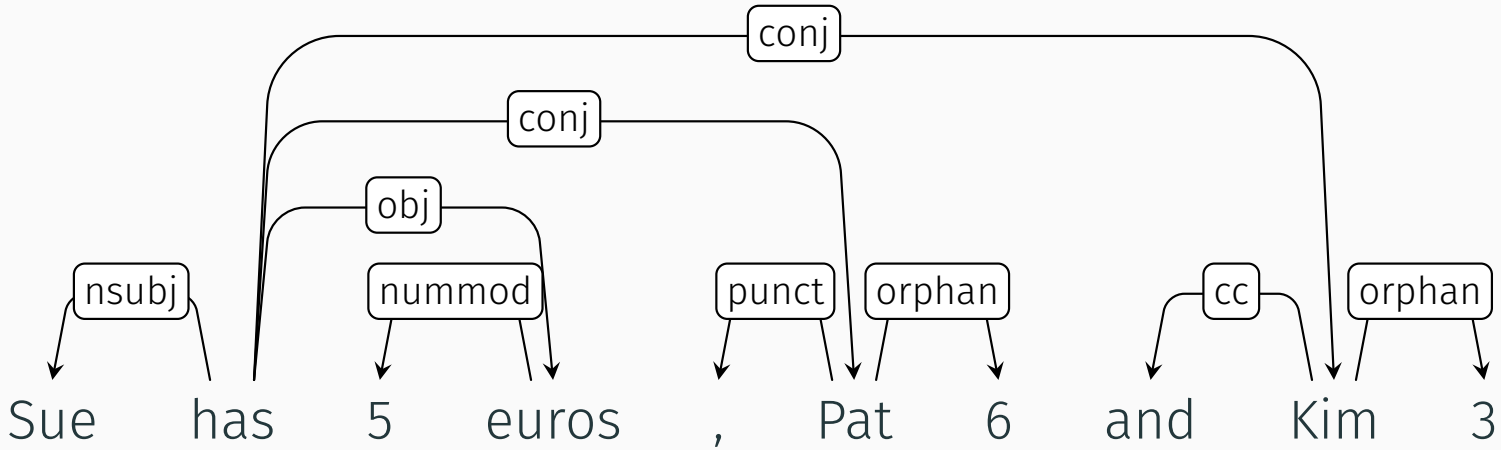


Enhanced UD



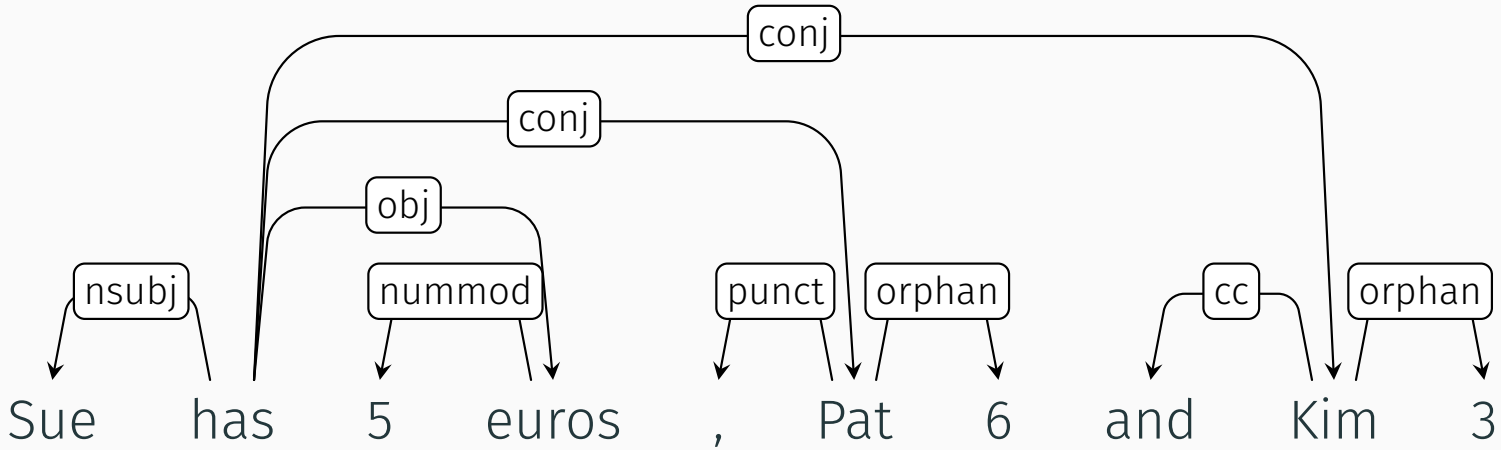
Gapping

Basic UD

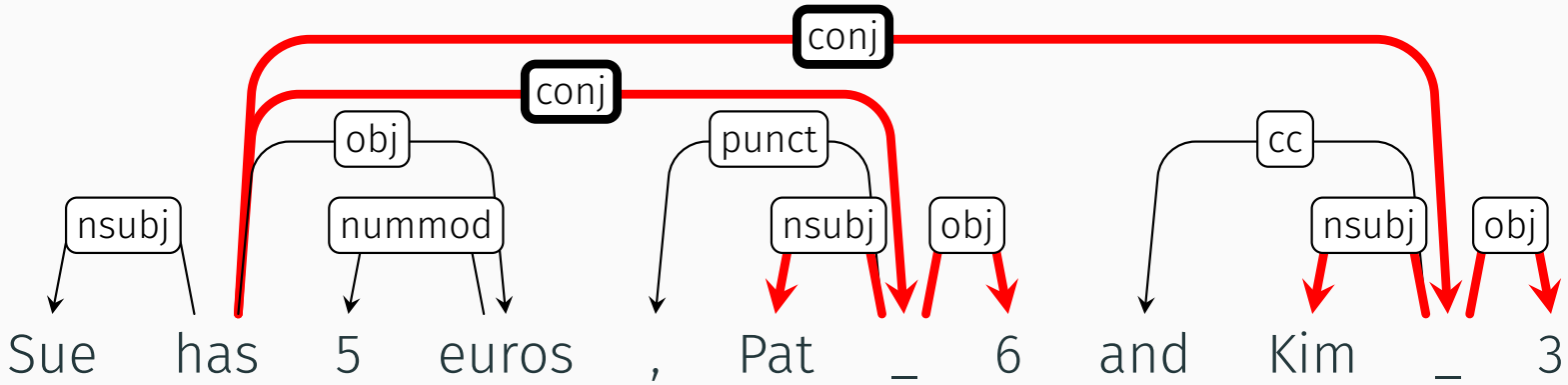


Gapping

Basic UD



Enhanced UD



IWPT Shared Task on Parsing into Enhanced UD

Motivation Are parsers able to produce enhanced UD annotation graphs automatically?

Data 17 languages (28 treebanks) in UD v2.5 contain Enhanced UD

Evaluation Labeled attachment F1-score over the set of Enhanced dependency triples (ELAS)

Approaches

- Output basic UD and convert to EUD
- Compile all of EUD into a basic UD-compatible format
- Parse into EUD directly (graph-based or transition-based)

Results of IWPT Shared Task

Team	LAS	ELAS
baseline	100.00	79.86
TurkuNLP	87.31	84.50
Orange	86.79	82.60
Emory NLP	86.14	79.84
FASTPARSE	77.57	74.04
UNIPI	80.74	72.76
ShanghaiTech	0.99	71.74
CLASP	82.66	67.85
ADAPT	84.09	67.23
Køpsala	75.41	62.91

- 'baseline' : copy gold standard UD into EUD
- Drop in going from UD to EUD is often less than 5%

Back to Linguistic Theory

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible

Back to Linguistic Theory

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible
- Enhanced UD is able to capture most of the information that is present in formal grammar models (HPSG, LFG)

Back to Linguistic Theory

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible
- Enhanced UD is able to capture most of the information that is present in formal grammar models (HPSG, LFG)
- Slightly different formats are possible as well:
  - Surface-Syntactic UD (Gerdes et al, 2017, 2019): capture (phrasal) distributional generalisations (function words are heads) repartitioning dependency relations, surface vs deep syntax

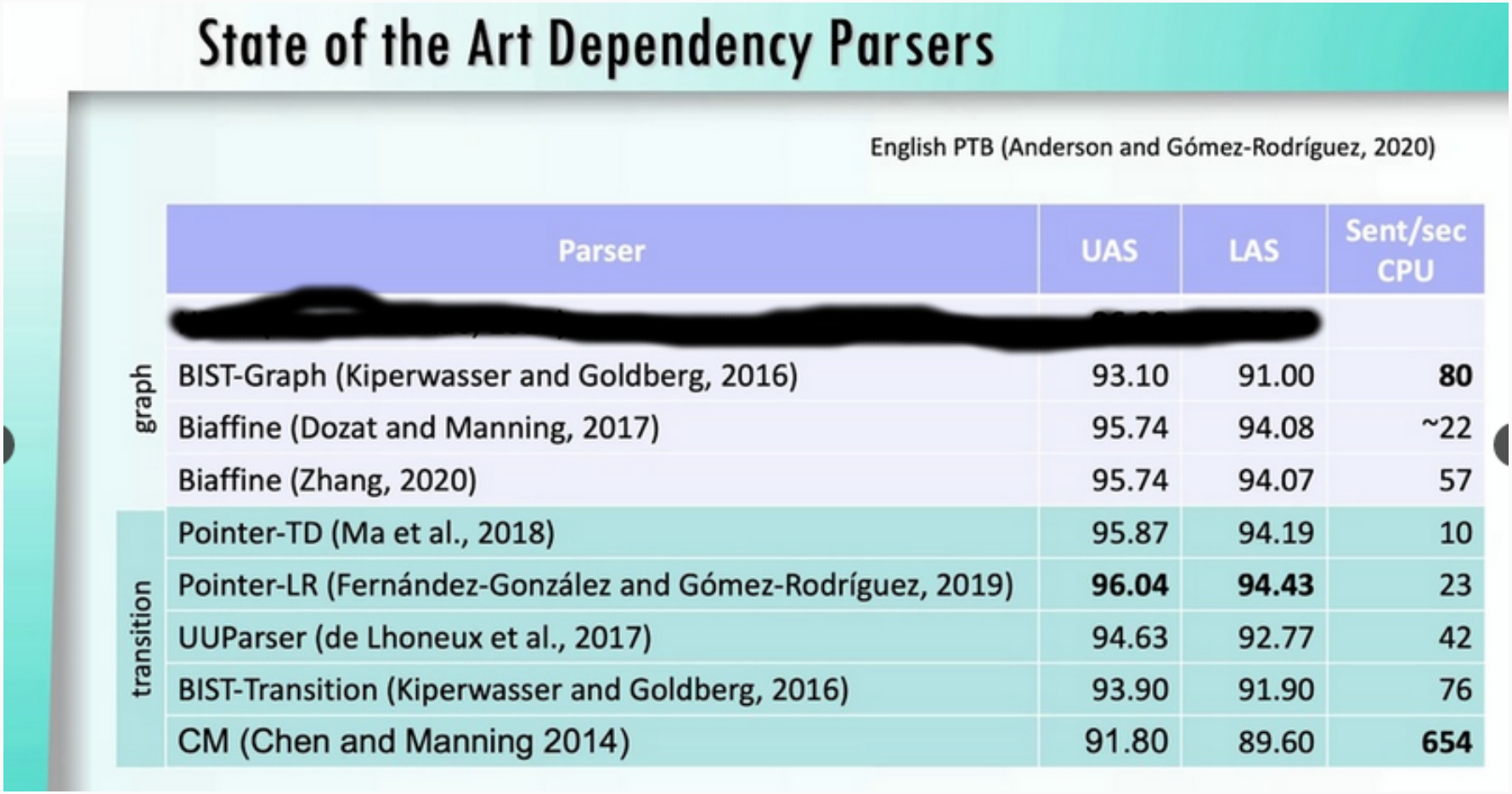
Back to Linguistic Theory

Enhanced UD vs Syntactic Frameworks

- Automatic parsing into Enhanced UD is feasible
- Enhanced UD is able to capture most of the information that is present in formal grammar models (HPSG, LFG)
- Slightly different formats are possible as well:
  - Surface-Syntactic UD (Gerdes et al, 2017, 2019): capture (phrasal) distributional generalisations (function words are heads) repartitioning dependency relations, surface vs deep syntax
- Semantic Interpretation is underdeveloped area:
  - Reddy, 2017, Gotham and Haug, 2018

Phrase Structure as a guide

Dependency Penn Treebank SOTA



State of the Art Dependency Parsers

English PTB (Anderson and Gómez-Rodríguez, 2020)

	Parser	UAS	LAS	Sent/sec CPU
	[redacted]			
graph	BIST-Graph (Kiperwasser and Goldberg, 2016)	93.10	91.00	**80**
graph	Biaffine (Dozat and Manning, 2017)	95.74	94.08	~22
graph	Biaffine (Zhang, 2020)	95.74	94.07	57
transition	Pointer-TD (Ma et al., 2018)	95.87	94.19	10
transition	Pointer-LR (Fernández-González and Gómez-Rodríguez, 2019)	**96.04**	**94.43**	23
transition	UUParser (de Lhoneux et al., 2017)	94.63	92.77	42
transition	BIST-Transition (Kiperwasser and Goldberg, 2016)	93.90	91.90	76
transition	CM (Chen and Manning 2014)	91.80	89.60	**654**

(*Linear Neural Parsing and Hybrid Enhancement for Enhanced Universal Dependencies*, Attardi *et al*, IWPT 2020)

Why are BERT models so successful?

The power of pre-trained Language Models

- LMs trained on large corpora already capture quite a bit a syntactic structure
- Therefore, limited amount of task-specific training data (i.e. treebank) already gives high accuracy

Why are BERT models so successful?

The power of pre-trained Language Models

- LMs trained on large corpora already capture quite a bit a syntactic structure
- Therefore, limited amount of task-specific training data (i.e. treebank) already gives high accuracy

Probing: Making syntactic predictions

- Linzen et al, 2016, Predict subject-verb agreement
  - The **keys** to the cabinet **are/*is** on the table
  - Alluvial **soils** carried in the floodwaters **add/*adds** nutrients to the floodplains.
  - Yet the **ratio** of men who survive to the women and children who survive **is/*are** not clear in this story.
- Also for relative clause attachment, negative polarity items, reflexives, relative pronouns (Dutch)

Why are BERT models so successful?

Probing for syntactic structure directly

- Hewitt and Manning, 2019: It is possible to learn a transformation of the vectorspace that predicts tree distance between words, as well as tree-depth of a word (i.e. no fine-tuning on task data)

Figure 3: Parse tree depth according to the gold tree (black, circle) and the norm probes (squared) on ELMo1 (red, triangle) and BERTLARGE16 (blue, square).

What about Phrase Strucure?

State of the Art Dependency Parsers

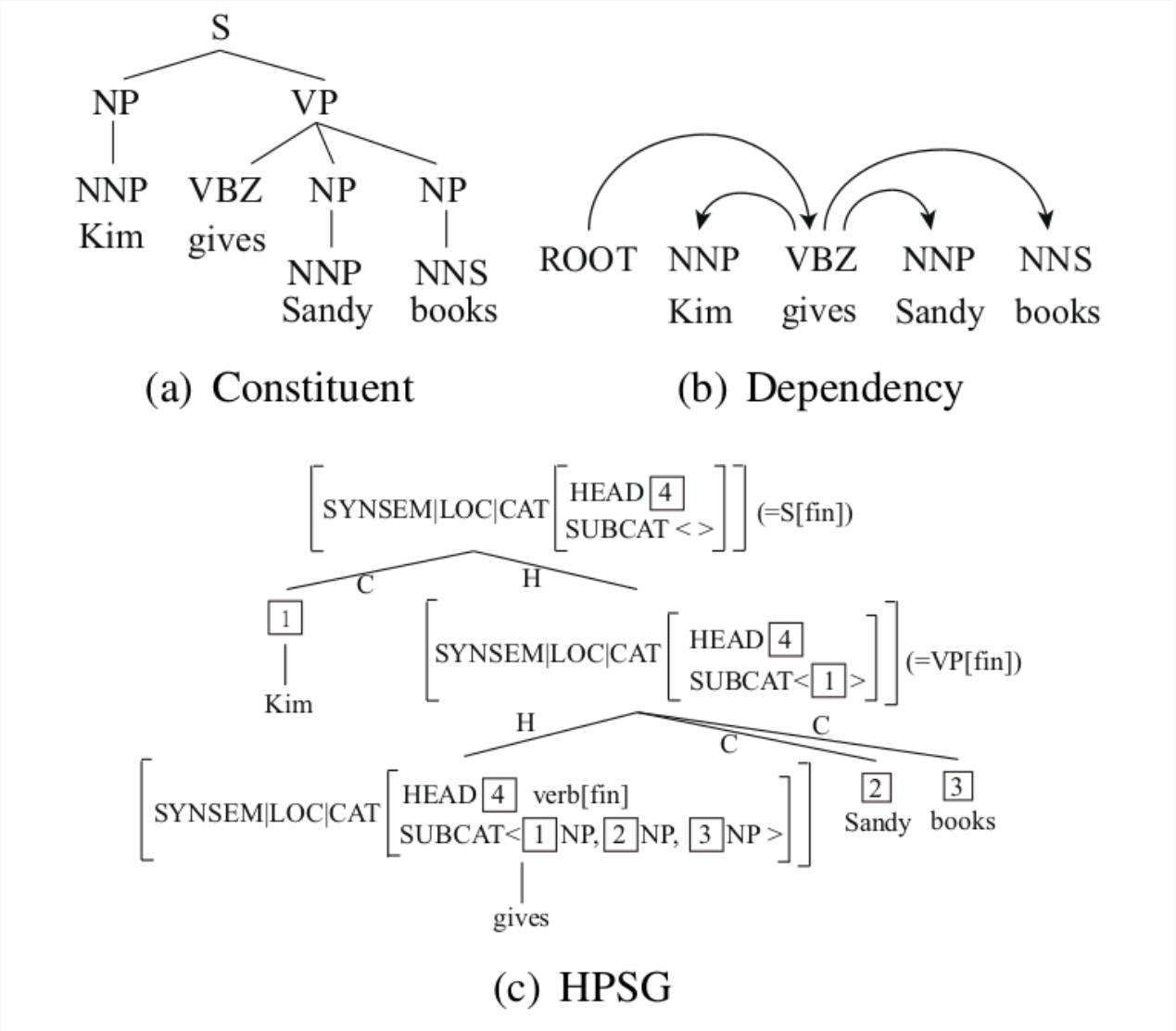
English PTB (Anderson and Gómez-Rodríguez, 2020)

	Parser	UAS	LAS	Sent/sec CPU
graph	HPSG (Zhou and Zhao, 2019)	**96.09**	**94.68**	
graph	BIST-Graph (Kiperwasser and Goldberg, 2016)	93.10	91.00	**80**
graph	Biaffine (Dozat and Manning, 2017)	95.74	94.08	~22
graph	Biaffine (Zhang, 2020)	95.74	94.07	57
transition	Pointer-TD (Ma et al., 2018)	95.87	94.19	10
transition	Pointer-LR (Fernández-González and Gómez-Rodríguez, 2019)	**96.04**	**94.43**	23
transition	UUParser (de Lhoneux et al., 2017)	94.63	92.77	42
transition	BIST-Transition (Kiperwasser and Goldberg, 2016)	93.90	91.90	76
transition	CM (Chen and Manning 2014)	91.80	89.60	**654**

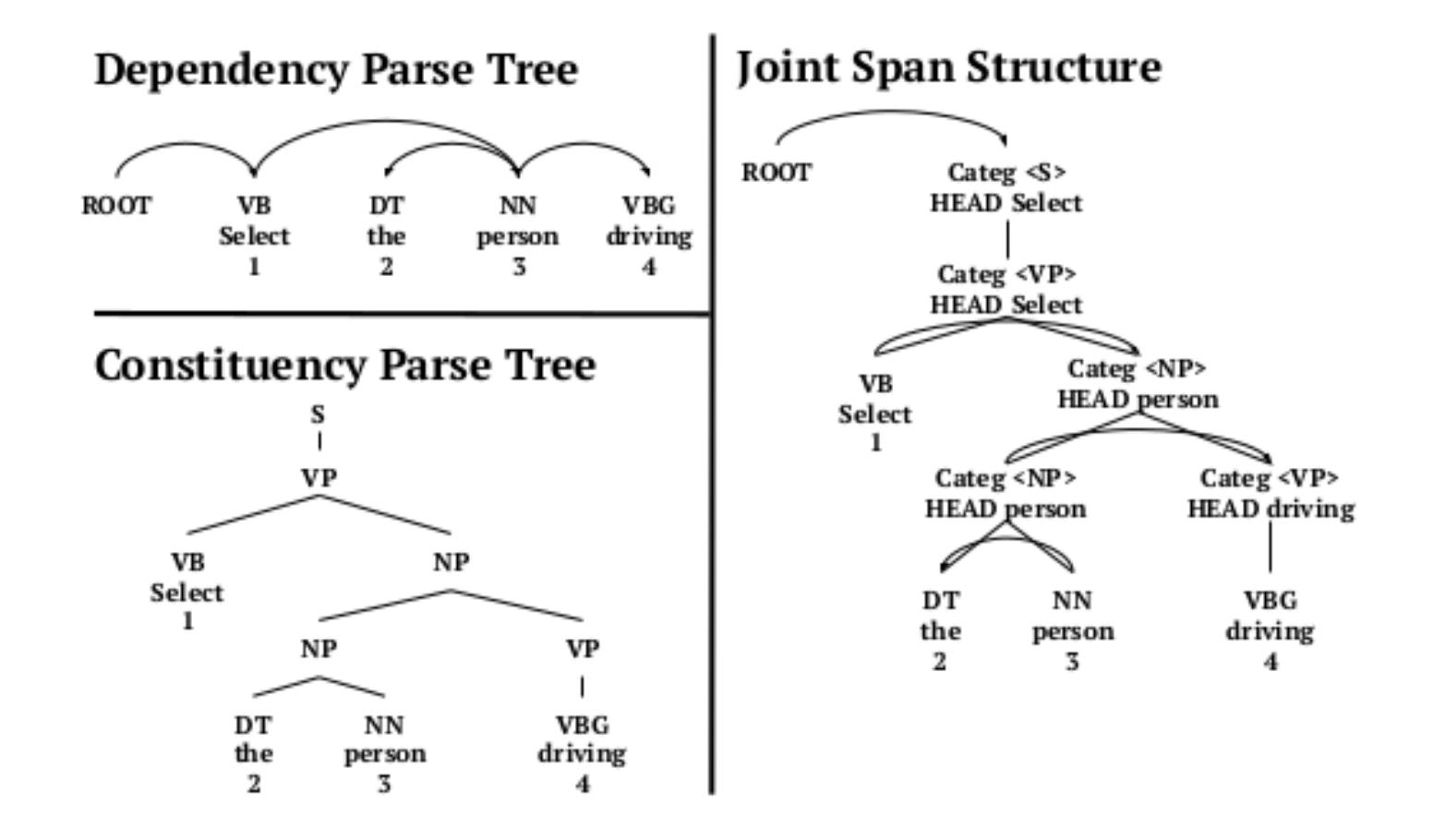
(*Linear Neural Parsing and Hybrid Enhancement for Enhanced Universal Dependencies*, Attardi *et al*, IWPT 2020)

Merging Phrase Structure and Dependencies

- Zhou and Zhao, 2019, Head-Driven Phrase Structure Grammar Parsing on Penn Treebank



Merging Phrase Structure and Dependencies



- Constituent parsing in the neural era:

  top down prediction of **splits** between spans

- *...our HPSG parsing model is more effective than learning constituent or dependency parsing separately*

Merging Phrase Structure and Dependencies

Constituency parsing

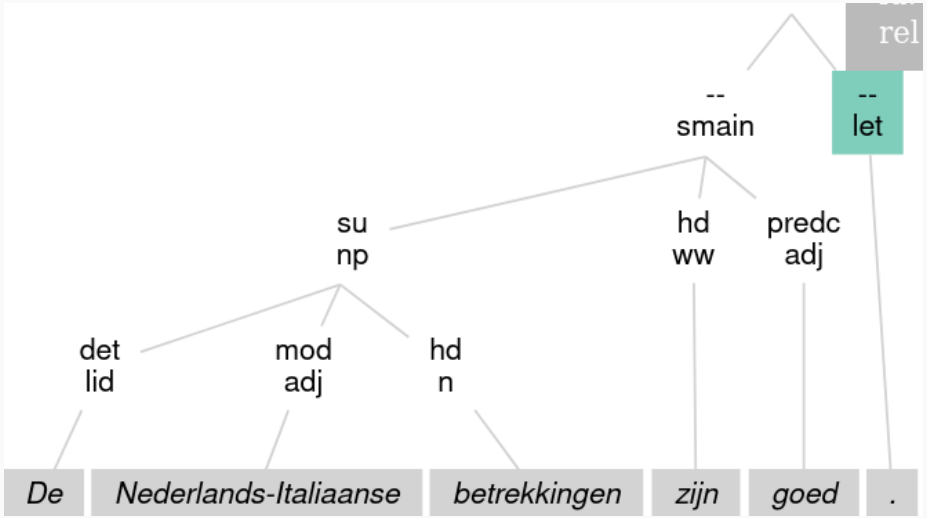
	EN F1	CH F1
Charniak 2016	93.8	
Fried & Klein 2018		87.0
Kitaev et al 2018	95.77	91.75
Zhao & Zhou 2019	96.33	92.18
Mrini et al 2020	**96.38**	**92.64**

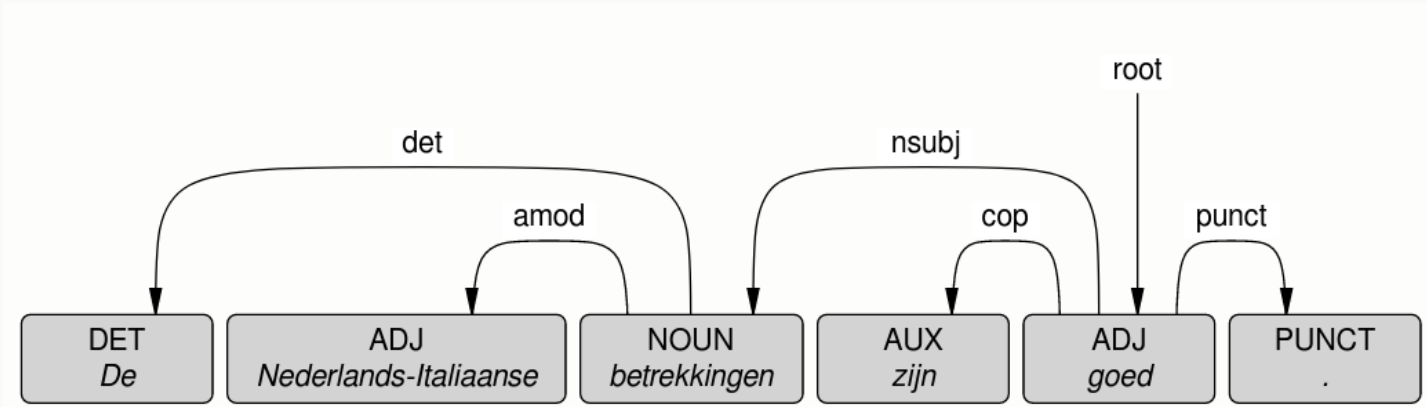
Dependency parsing

	EN LAS	CH LAS
Dozat and Manning 2016	94.08	88.23
Zhao & Zhou 2019	95.72	89.15
Mrini et al 2020	**96.26**	**89.26**

Mrini et al, 2020 Rethinking self-attention: towards interpretability in neural parsing

Can we apply this trick for Dutch?

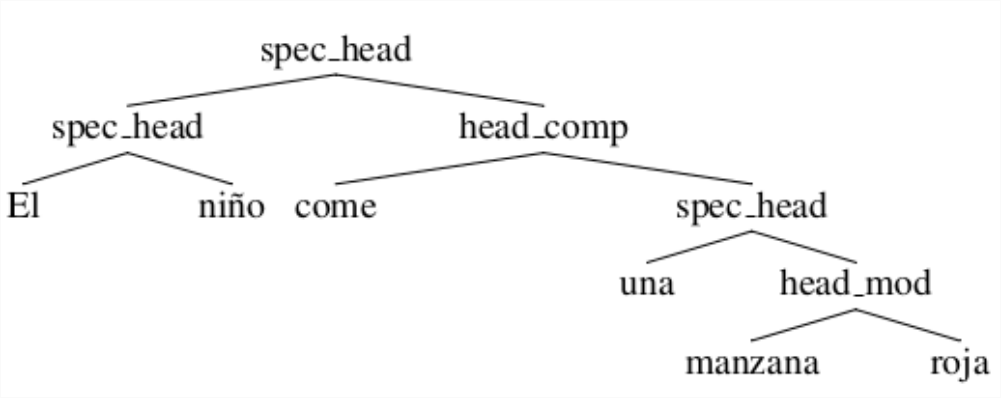




The Dutch-Italian relationships are good.

Can we apply this trick more widely?

Spanish Ancora corpus

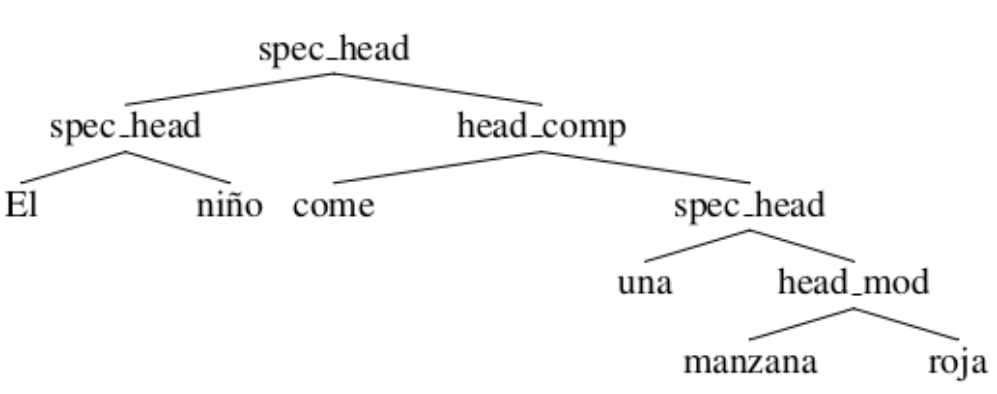


the boy eats a red apple

Chiruzzo and Wonserver, 2020,Statistical Deep Parsing for Spanish

Can we apply this trick more widely?

Spanish Ancora corpus



the boy eats a red apple

Chiruzzo and Wonserver, 2020,Statistical Deep Parsing for Spanish

UD v2.5

- 158 treebanks, 100 treebanks are converted from non-UD data:

`Relations | annotated manually in non-UD style, automatically converted to UD`

Conclusions

(Enhanced) Universal Dependencies and Phrase Structure

- Enhanced UD and related formats capture substantial portion of information expressed by formal syntactic theories
- HPSG-inspired combinations of dependency annotation and phrase structure contribute to accurate parsing

Questions?

Tuesday, August, 18, 15:25-15:45
(Paris/Berlin/Amsterdam time)