Quantifying training challenges of dependency parsers

Lauriane Aufrant[1,2], *Guillaume Wisniewski*[1], *François Yvon*[1]

(1) LIMSI, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, (2) DGA

{lauriane.aufrant, guillaume.wisniewski, francois.yvon}@limsi.fr

Overview

Is parsing an easy task?

↪ **10 training sentences** often suffice for correctly predicting **70% of dependencies**

Yet, those are **learned in very different ways**:

- DET, AUX: straightforward rules
  simple: require few examples
- VERB: semantics-driven
- mwe: enumeration-based
  complex: require many examples

Contributions:

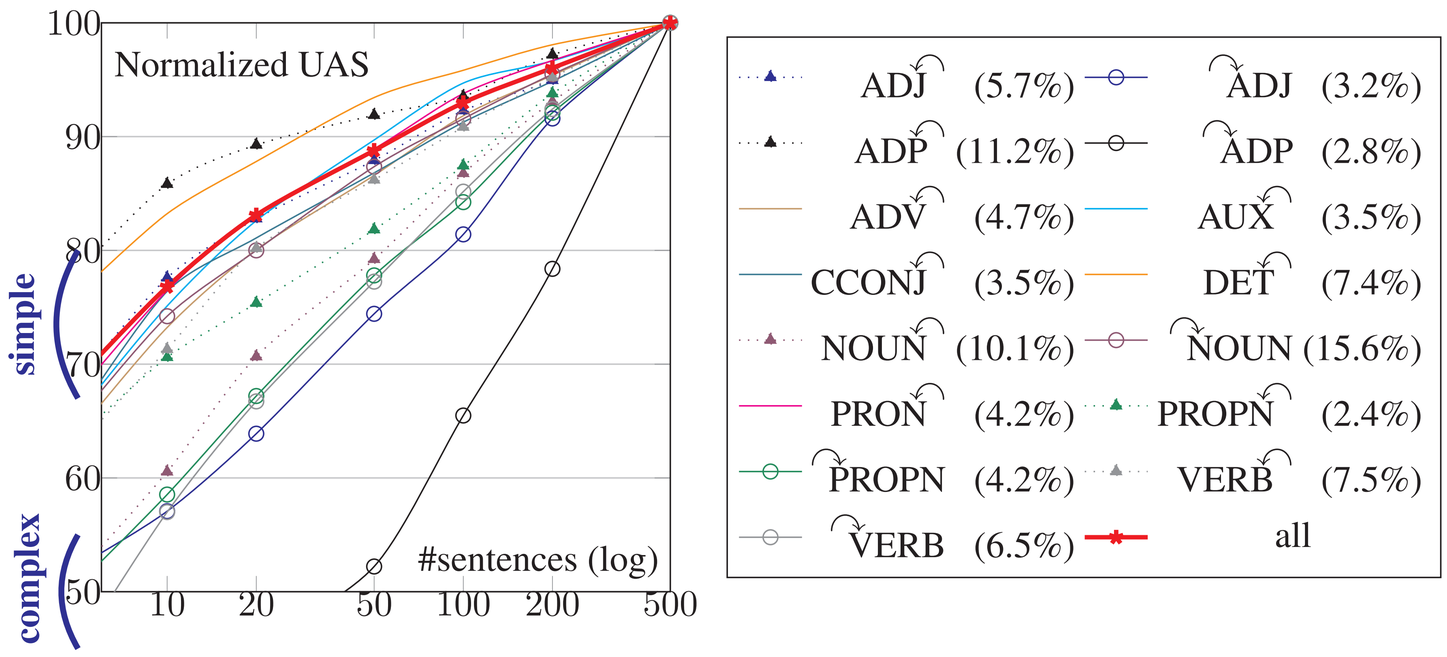
↪ Formalization of a new property for dependencies

↪ New tools for fine-grained analysis of parsers

Class complexity = area between the average and class-level curves (in log scale)

Class-level learning rate [perceptron-based Beam parser, UD 2.0]

Custom visualization: the slope represents the **marginal utility** of **doubling** the treebank size



Large-margin grouping into **2 categories**:

↪ simple curves quickly saturate

↪ complex classes better leverage additional data

Class frequencies explain **only partially** the ordering:

↪ DET vs ADJ (but cf NOUN)

↪ ADJ vs ADJ (but cf PROPNs)

Complexity measures [perceptron-based Beam parser, UD 2.0]

Language-independent trends:

Complexity	ADP	DET	PRON	AUX	ADJ	CCONJ	N	ADV	V	PN	SCONJ	N	PN	V	ADJ	AUX	ADP
	-18.8	-18.7	-0.6	0.2	1.9	6.3	7.6	9.6	12.6	23.4	35.0	42.0	49.5	52.5	57.7	68.0	131.2
UAS_{500}	DET	ADP	AUX	PRON	SCONJ	ADJ	CCONJ	ADV	V	PN	PN	N	N	ADJ	V	AUX	ADP
	91.3	89.0	83.9	82.4	80.2	80.0	77.1	76.1	75.1	69.0	68.4	68.2	67.9	60.6	56.4	52.8	48.0

Language-level variations:

- ADJ and ADJ are simple/complex in English, complex/simple in French
- ADP is usually complex (50 treebanks out of 56), DET is usually simple (49 treebanks out of 56)

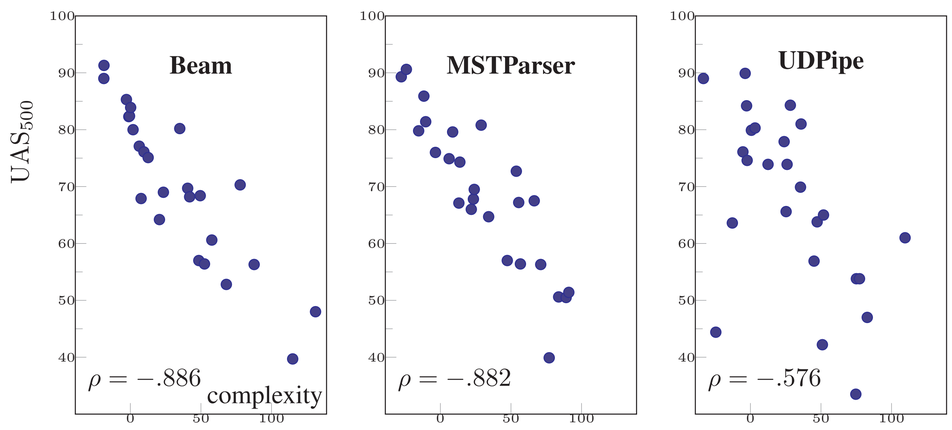
Application 1: Enabling deeper analyses of parsing results

Identify shortcomings on specific classes:

↪ hints at **parser properties**

- For Beam, AUX are simpler
- For MSTParser, VERB are simpler — **thanks to non-local features?**
- For UDPipe, CCONJs are simpler, DET are less accurate while being less simple — **is their determinism under-exploited?**

Class-level score/complexity correlation:



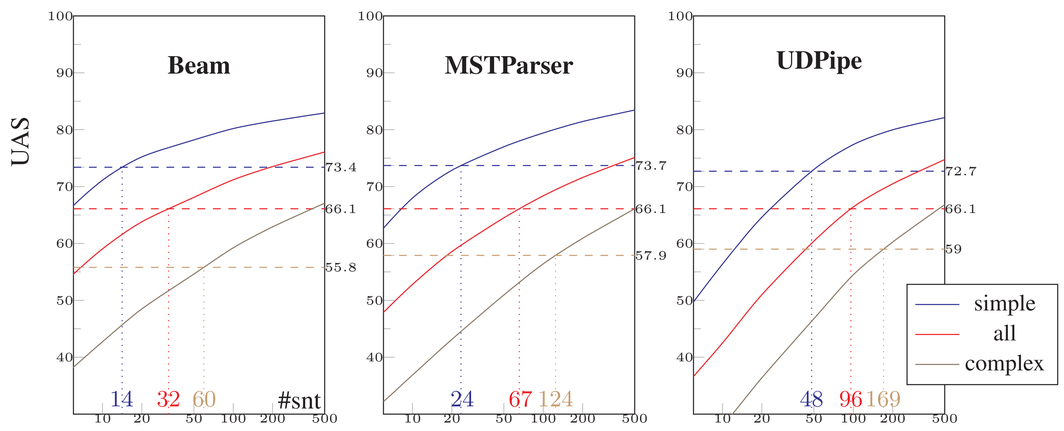
Legible composite scores:

	UAS_{10}			UAS_{500}			$UAS_{\text{full UD}}$		
	simple	overall	complex	simple	overall	complex	simple	overall	complex
UDPipe	52.8	42.5	28.4	81.8	74.7	65.2	87.6	83.2	**77.3**
MSTParser	64.4	52.8	36.9	82.7	75.1	64.9	**88.2**	**83.4**	77.1
Beam	**71.1**	**59.0**	**42.7**	**82.9**	**76.1**	**67.1**	87.3	82.6	76.4

- Significant **score gap** maintained between both categories, even for large datasets
  ↪ confirms that they capture different properties
- Reveals the **strengths** of each parser, and how they **leverage additional data**

Application 2: What do cross-lingual parsers learn?

Using multi-source weighted delexicalized transfer:



- On average, as accurate as **training on 32 sentences** (using Beam)
- But qualitatively **more informative** (better on complex classes)

A wide array of possible applications:

↪ assessing methods for annotation projection, unsupervised parsing, domain adaptation...

Case study: information conveyed by one source

Sources	UAS (ro)			#sentences (ro)		
	simple	overall	complex	simple	overall	complex
fr + it + es + bg	81.4	74.4	60.3	167	213	231
✗ it es bg	81.2	73.8	59.1	149	165	179
fr ✗ es bg	81.1	73.6	58.5	142	155	162
fr it ✗ bg	79.9	73.0	59.1	77	131	179
fr it es ✗	79.9	73.3	60.1	77	142	219

↪ **Added value** of each source:

- Italian, French: information on complex classes
- Bulgarian: information on simple classes (not provided by Romance sources)
- Spanish: both

↪ Italian is a **qualitatively better** source than Spanish (despite similar scores)

↪ **Complementarity** of Spanish and Bulgarian: **doubles** the simple treebank size

COLING'18 – International Conference on Computational Linguistics