

# Annotating Wall Street Journal Texts Using a Hand-Crafted Deep Linguistic Grammar



Valia Kordoni & Yi Zhang LT-Lab, German Research Center for Artificial Intelligence (DFKI GmbH) & Dept. of Computational Linguistics, Saarland University, Germany {kordoni,yzhang}@coli.uni-sb.de

[0]

[1]

[2]

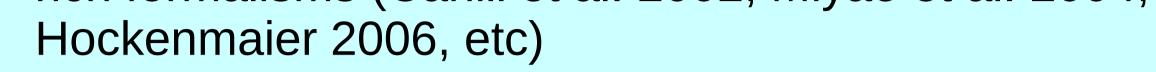
[3]

### Introduction

- An on-going project to annotate texts from WSJ sections of PTB (~1M tokens, ~43K utterances) with HPSG analyses
- Independent annotation from PTB by parsing with ERG and manual disambiguation of the outputs
- Annotation only for sentences covered by the grammar
- Different to conversion-based development of treebanks in rich formalisms (Cahill et al. 2002, Miyao et al. 2004,

## Treebanking

- For each tree, up to 500 candidates (ranked by a statistical disambiguation model) are recorded
- Discriminant-based treebanking decisions
- Yes/No choice for candidate discriminants
- For n trees, on average need  $\log_2 n$  steps to fully disambiguate



•  $\approx$  dynamic treebank (Oepen et al. 2002)

#### **HPSG** Parsing

- English Resource Grammar (Flickinger 2002)
- PET Parser (Callmeier 2001)
- [incr tsdb()] (Oepen & Callmeier 2000)
- Preprocessing (Adolphs et al. 2008)
- Robust Parsing \* (Zhang et al. 2008)

#### Annotation Cycle

#### **Grammar & Treebank Update**

• Update of the disambiguation model:

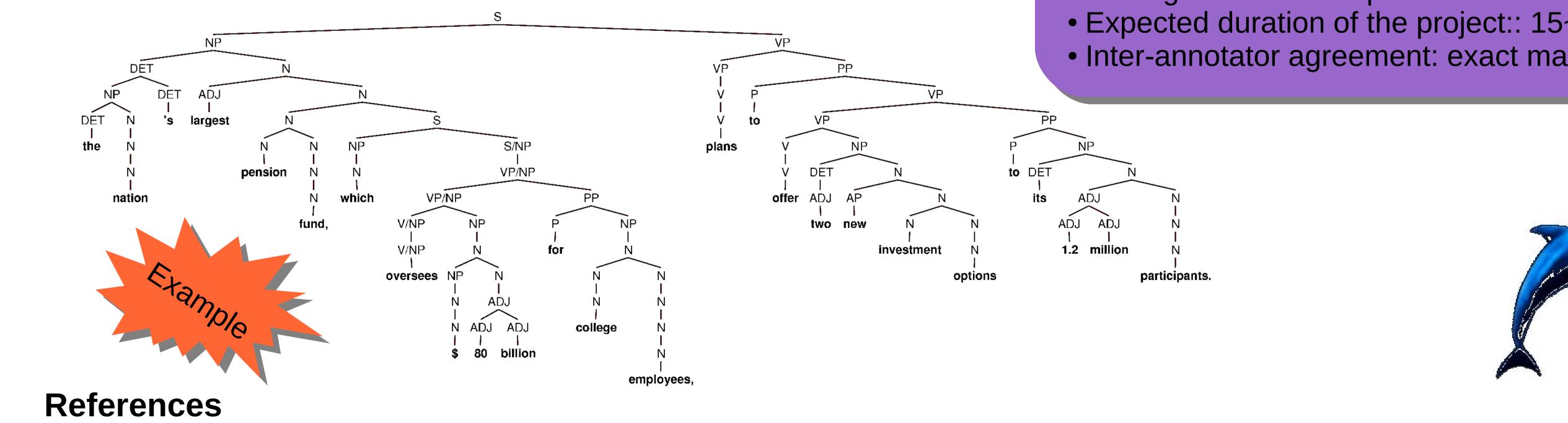
#### Close Previous Next Reject Clear Reset Ordered Concise Full Save Confidence Toggle

(20200002) [15 : 0 @ high] ((S (NP-SBJ (NNP Ms.) (NNP Haag)) (VP (VBZ plays) (NP (NNP Elianti))) (. .)))

S	A	Ms. Haad plavs Elianti .				
XP S				1		
NP NP VP	2	??	APPOS	ms. Haag plays    Elianti.	$\Delta$	
	12	/ ? តែគា	FRAG_NP	ms. Haag plays Elianti.		
		] []	SUBJH	ms. Haag    plays Elianti.	Yes	
Ň Ň Ý Ň			RUNON_S	ms.    Haag plays Elianti.	No	
Ń Haag plays Ń		2 2	NP_N_CMPND NOUN_N_CMPND	ms. Haag    plays ms.    Haag plays	Unknown	
ms. Elianti.	$  _{2}$	2 2	BARE_NP	ms. Haag plays	In Parses (4)	
S	2	27	APPOS	Haag plays    Elianti.	Out Parses (11)	
	2		FRAG_NP	Haag plays Elianti.		
XP S	2	2	SUBJH	Haag    plays Elianti.		
NP NP VP	2	??	APPOS	ms.    Haag		
	?	??	NP_TITLE_CMPND	ms.    Haag		
	?	??	NP_N_CMPND	Haag    plays		
	?	??	BARE_NP	Haag plays		
N Haag plays N	?	??	HCOMP	plays    Elianti.		
ms. N	?	??	SING_NOUN_IRULE	ms.		
Elianti.	2		PROPER_NP	ms.		
S	2	??	FRAG_NBAR	ms.		
NP VP	2	2 2	MASS_COUNT_IRULE PUNCT_PERIOD_ORULE	ms. ms.		
		27	BARE_NP	ms.		
	2		FRAG_NP	ms.		
	2	2.7	npn-abb_le	ms.		
ms. N plays N	2	??	nmc_le	ms.		
Haiag N	?	??	SING_NOUN_IRULE	Haag		
Elianti.	?		PLUR_NOUN_ORULE	plays		
S			THIRD_SG_FIN_VERB_ORULE			
			nmc_le	plays		
			v_np*_le	plays Fliortí		
	1		PUNCT_PERIOD_ORULE	Elianti.		
N NP NP					-	
		21			IN IN	

#### **Quality Assurance**

- newly annotated sections are used to train better parse selection models that help annotators to find the best tree more quickly
- Update of the Grammar: grammar writer receives feedback from annotators and fixes systematic errors or improves coverage
- Update of the annotation: existing annotations will be semi-automatically synchronized with the new versions of the grammar, with some extra manual annotation required



- Multiple annotators with partially overlapping assignments. Inter-annotator agreement is checked regularly
- Regular treebankers' meetings to discuss issues in treebanking and annotation strategies to achieve higher inter-annotator agreement

#### **Some Facts in Numbers**

- Number of annotators: 3
- Grammar parsing coverage: ~80%
- Average annotation speed: 35~40 sentences/hour
- Expected duration of the project:: 15~18 month
- Inter-annotator agreement: exact match agreement of ~50%



- Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank: motivation and preliminary applications. In *Proceedings of COLING 2002*. Taipei, Taiwan.
- Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In Collaborative Language Engineering, pages 1–17. CSLI Publications.
- Yi Zhang and Valia Kordoni. 2008. Robust Parsing with a Large HPSG Grammar. In Proceedings of LREC 2008, Marrakesh, Morocco.
- Ulrich Callmeier. 2001. *Efficient parsing with large-scale unification grammars*. Master's thesis, Saarland University, Saarbruecken, Germany.
- Stephan Oepen and Ulrich Callmeier. 2000. Measure for Measure: Parser Cross-Fertilization. Towards Increased Component Comparability and Exchange. In Proceedings of the 6th International Workshop on Parsing Technology, Trento, Italy.
- Peter Adolphs, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Daniel Flickinger, Bernd Kiefer. 2008. Some Fine Points of Hybrid Natural Language Parsing. In *Proceedings of LREC 2008*, Marrakesh, Morocco.