CCG grammar extraction from treebanks: translation algorithms and applications

Julia Hockenmaier

(and Mark Steedman, Johan Bos, Ruken Cakici, Stephen Clark, James Curran, Dan Gildea, Mike White,)

http://www.cis.upenn.edu/~juliahr
http://groups.inf.ed.ac.uk/ccg

Three languages & corpora

Three languages & corpora

	English	German	Turkish

Three languages & corpora

	English	German	Turkish
Source corpus	**Penn Treebank:** phrase structure grammar	**Tiger Corpus:** “syntax graphs”	**METU-Sabancı:** dependencies

Three languages & corpora

	English	German	Turkish
Source corpus	**Penn Treebank:** phrase structure grammar	**Tiger Corpus:** “syntax graphs”	**METU-Sabancı:** dependencies
Output corpus	**Derivations & dependencies** Hockenmaier & Steedman (2002, 2005, 2007)	**Currently just derivations** Hockenmaier (2006)	**Lexical categories** Cakici (2005,2007)

Three languages & corpora

	English	German	Turkish
Source corpus	**Penn Treebank:** phrase structure grammar	**Tiger Corpus:** “syntax graphs”	**METU-Sabancı:** dependencies
Output corpus	**Derivations & dependencies** Hockenmaier & Steedman (2002, 2005, 2007)	**Currently just derivations** Hockenmaier (2006)	**Lexical categories** Cakici (2005,2007)
Binary trees	Flat trees & head rules	Non-planar graphs with heads	?

Three languages & corpora

	English	German	Turkish
Source corpus	**Penn Treebank:** phrase structure grammar	**Tiger Corpus:** “syntax graphs”	**METU-Sabancı:** dependencies
Output corpus	**Derivations & dependencies** Hockenmaier & Steedman (2002, 2005, 2007)	**Currently just derivations** Hockenmaier (2006)	**Lexical categories** Cakici (2005,2007)
Binary trees	Flat trees & head rules	Non-planar graphs with heads	?
Comp/adj distinction	Function tags	Edge labels	given

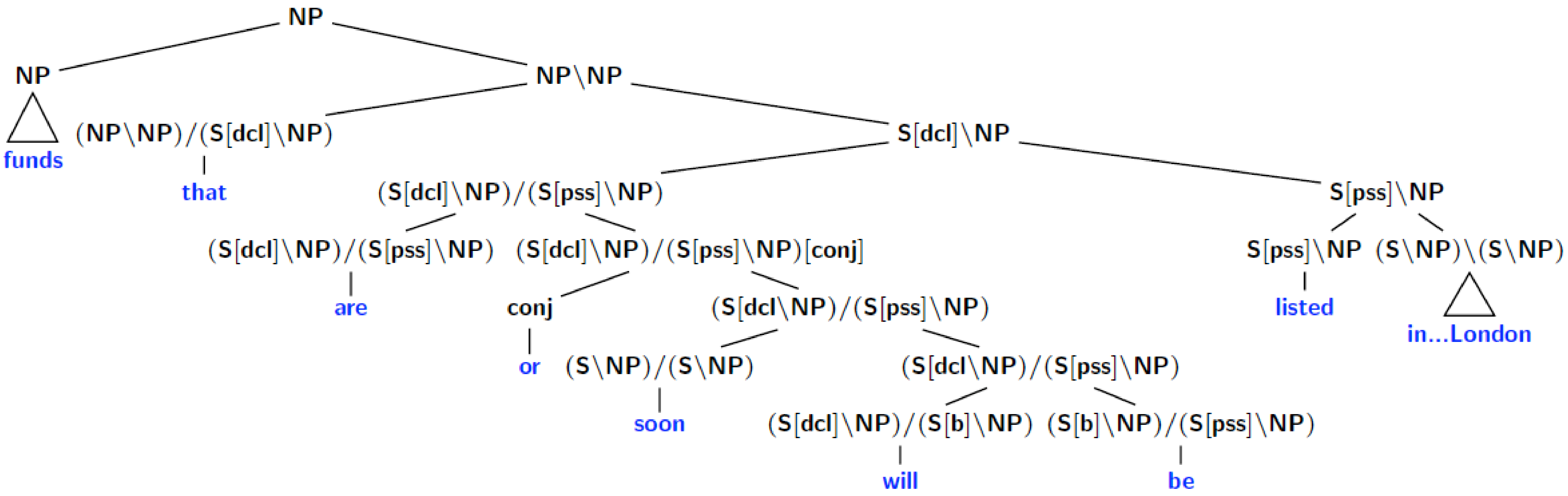
Three languages & corpora

	English	**German**	**Turkish**
Source corpus	**Penn Treebank:** phrase structure grammar	**Tiger Corpus:** “syntax graphs”	**METU-Sabancı:** dependencies
Output corpus	**Derivations & dependencies** Hockenmaier & Steedman (2002, 2005, 2007)	**Currently just derivations** Hockenmaier (2006)	**Lexical categories** Cakici (2005,2007)
Binary trees	Flat trees & head rules	Non-planar graphs with heads	?
Comp/adj distinction	Function tags	Edge labels	given
Unbounded dependencies	Traces & null elements	Secondary edges	Requires manual reannotation

What CCGbank encodes

- **Syntactic categories/derivations:**
  - **Derivations are binary trees**
  - **Categories encode functor-argument relations**
    (head-complement or modifier-head)
  - **Lexical categories = subcat frames**
  - **Unbounded non-local dependencies**
    wh-movement, right-node raising, argument cluster coordination
  - **Bounded non-local dependencies** (raising, control)
  - **Syntactic categories correspond to semantic types!**

- **Word-word dependency structures:**
  - ***Non-anaphoric* local and non-local dependencies**



that	**((NP\NP)/(S[dcl]\NP))**	funds	are, will	
are	**((S[dcl]\NP)/(S[pss]\NP))**	*funds*	*listed*	
soon	**((S\NP)/(S\NP))**		will	
will	**((S[dcl]\NP)/(S[b]\NP))**	*funds*	be	
be	**((S[b]\NP)/(S[pss]\NP))**		listed	
listed	**(S[pss]\NP)**	*funds*		
in	**(((S\NP)\(S\NP))/NP)**		listed	York, London

Hockenmaier & Steedman, *Computational Linguistics* 33(3)

The need for preprocessing

- **Cleaning up noise:**
  - ✓ POS tagging errors
    (required for head-finding, features on categories)
- **Adding linguistic structure:**
  - ✓ Detecting coordination
  - ✓ Analyzing FRAGs, QPs, parentheticals
- **Changing linguistic analyses:**
  - ✓ Small clauses

Remaining problems

- **At the VP level:**
  - Complement/adjunct distinction
  - Phrasal verbs, particle-verb constructions
  - Heavy NP shift
- **At the NP level:**
  - Compound nouns
  - Coordinate nouns
  - Appositives vs. lists
  - Lack of number agreement
  - Attachment of NP modifiers

Problems arising in applications

- **Translation to DRS (e.g. for textual entailment)**
  Bos et al. (2004), Bos (2005),
  - Problems with NPs:
    quantifying NPs, restrictive rel. clauses, compound nouns
- **Semantic role labeling**
  Gildea and Hockenmaier (2003)
  - Problems with VPs (mismatches with Propbank)
    modifier scope, argument/adjunct distinction

Implications for treebank design

- **Some postprocessing is inevitable:**

  Linguistic analyses differ.

  But -- cleaning up noise is too expensive.
- **Explicit, detailed information matters:**

  Manually adding information is expensive.

Theories impose constraints on annotations, but minimal requirements are not formalism-specific!

- Heads, arguments, modifiers, conjuncts
- Non-local bounded and unbounded dependencies
- Distinction between different types of dependencies

But formalism-neutral annotation might be better:

- Annotation is description. Cheaper than theory-based analysis?
- Theories change, and might not account for data.