Update on Treebanking the WSJ with the ERG

Dan Flickinger, Yi Zhang, Valia Kordoni

DELPH-IN Summit, Sofia

July 3, 2012

Goals

- **Annotate** the Penn Treebank WSJ corpus with ERG analyses
- **Distribute** this treebank in a variety of forms

  Derivation trees

  Labeled bracketings

  MRS (and perhaps dependencies)
- **Use** as springboard for other research efforts

  Training better PCFGs

  Sharpening focus for robust processing

  Providing illustrations of linguistic phenomena

Motivation

- Common ground for evaluation and comparison
- Linguistically challenging but parsable text
- Rich array of other annotation resources available

Methodology

- Parse the 50,000 sentences using ERG + PET (but not Sec. 23)
- Manually disambiguate each parsed sentence (Redwoods)
- On second pass, update treebank using newest version of ERG
- Do quality control over treebank to detect errors
- On third pass, correct errors, and update using stable ERG

Methodology

- Parse the 50,000 sentences using ERG + PET (but not Sec. 23)

  Initially with POS-tagged structures, which we preprocess

  Probably change to parsing original text
- Manually disambiguate each parsed sentence (Redwoods)
- On second pass, update treebank using newest version of ERG
- Do quality control over treebank to detect errors
- On third pass, correct errors, and update using stable ERG

Methodology

- Parse the 50,000 sentences using ERG + PET (but not Sec. 23)

  Initially with POS-tagged structures, which we preprocess

  Probably change to parsing original text
- Manually disambiguate each parsed sentence (Redwoods)

  Multiple annotators, but mostly one annotator per sentence

  Accumulate shortcomings/inconsistencies in ERG en route
- On second pass, update treebank using newest version of ERG
- Do quality control over treebank to detect errors
- On third pass, correct errors, and update using stable ERG

Methodology

- Parse the 50,000 sentences using ERG + PET (but not Sec. 23)
  Initially with POS-tagged structures, which we preprocess
  Probably change to parsing original text
- Manually disambiguate each parsed sentence (Redwoods)
  Multiple annotators, but mostly one annotator per sentence
  Accumulate shortcomings/inconsistencies in ERG en route
- On second pass, update treebank using newest version of ERG
  Single annotator (distinct from those in first pass)
  Train WSJ-specific maxent model for use on remainder
- Do quality control over treebank to detect errors
- On third pass, correct errors, and update using stable ERG

Methodology

- Parse the 50,000 sentences using ERG + PET (but not Sec. 23)

  Initially with POS-tagged structures, which we preprocess

  Probably change to parsing original text

- Manually disambiguate each parsed sentence (Redwoods)

  Multiple annotators, but mostly one annotator per sentence

  Accumulate shortcomings/inconsistencies in ERG en route

- On second pass, update treebank using newest version of ERG

  Single annotator (distinct from those in first pass)

  Train WSJ-specific maxent model for use on remainder

- Do quality control over treebank to detect errors

  Distribute in three stages DELPH-IN-internally

- On third pass, correct errors, and update using stable ERG

Status and Schedule

- First pass: now almost complete annotations for all WSJ sections
- Second pass: one-third now ready for DELPH-IN-internal release
- Third (final) pass: aiming for completion by end of 2012
- Will submit paper for 11th TLT in Lisbon in December
- Announce public availability of ‘full’ treebank in early 2013

Some numbers on the first third

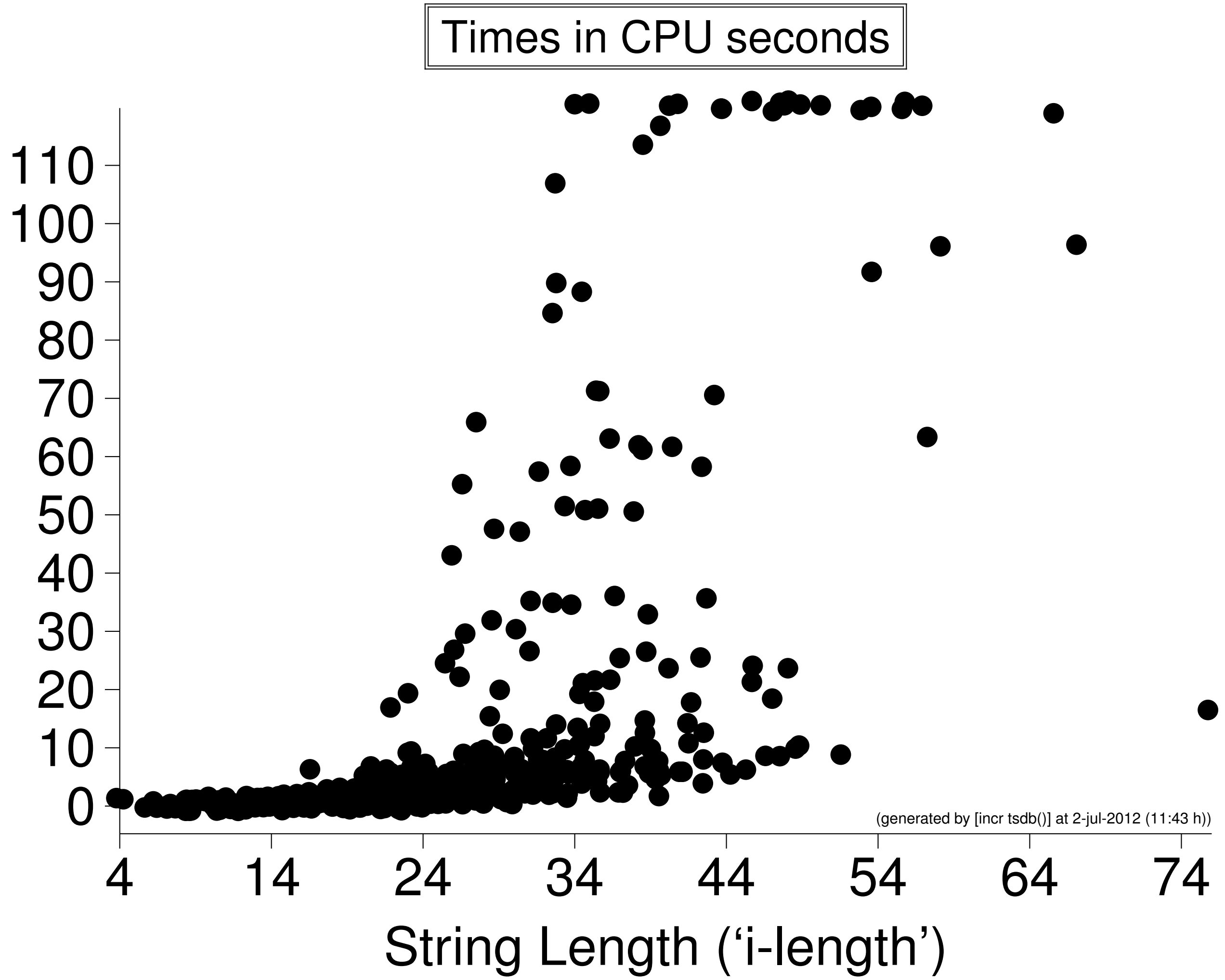
WSJ Items	Parsed	Banked
14,881	13,640	12,006
	91.7%	80.7%

- Average tokens per sentence: 25.4
- Resource limits per sentence for parsing:

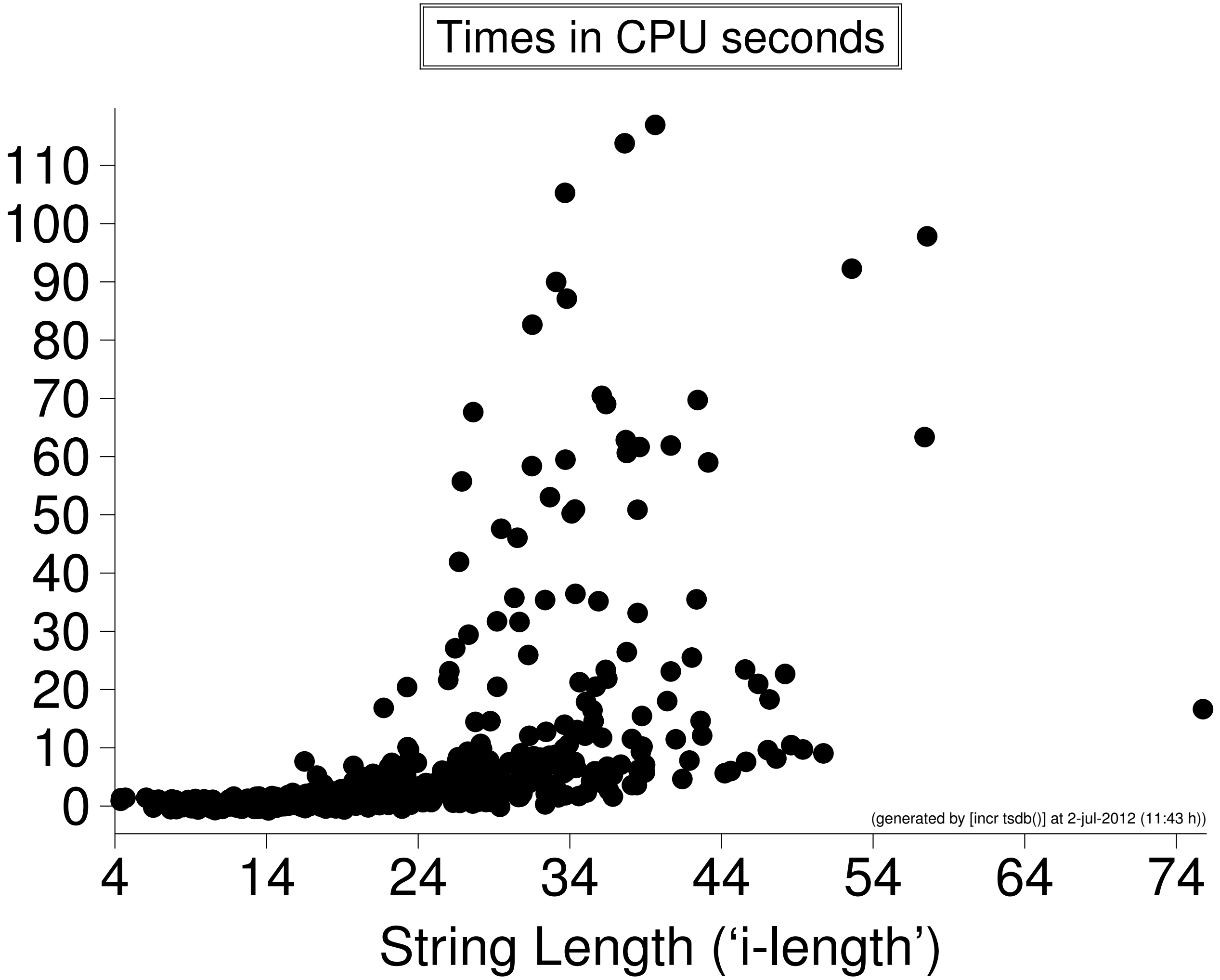
  120 CPU seconds

  200,000 edges

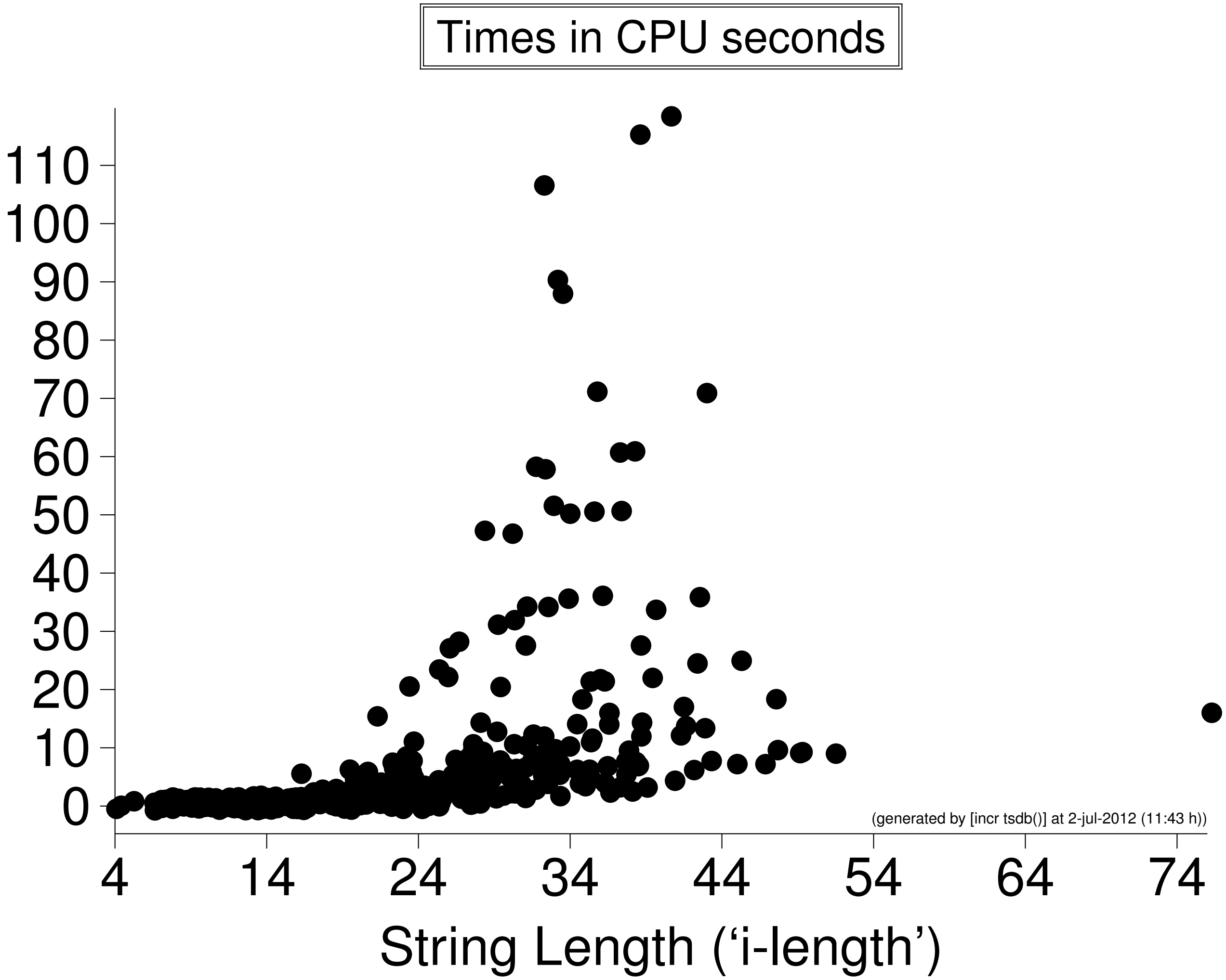
Parsing times: WSJ07d (all items)



Parsing times: WSJ07d (parsed)



Parsing times: WSJ07d (banked)



Internal Release to DELPH-IN

- Full and one-best profiles in SVN hosted at CoLi Saarbrücken

  As of now, sections 02-10a available

  If interested, contact Yi for access; please do not redistribute
- Feedback welcome – contact Dan

  Item-specific errors

  Systematic infelicities (syntax or semantics)

Acknowledgements

- We are grateful for Erasmus Mundus support to CoLi Saarbrücken
- Thanks to CoLi annotators, especially Iliana Simova

Naming the Resource

Naming the Resource

The Urban Forest

Naming the Resource

Central Park

Naming the Resource

ERG RoTree-bank

Naming the Resource

Ergonymic Forestree

Naming the Resource

Pulp Fiction

Naming the Resource

Hardwood Treebank