From the Jungle to a Park: Harmonizing Dependency Treebanks of 30 Languages

Coordination styles and transformations

Martin Popel
ÚFAL (Institute of Formal and Applied Linguistics)
Charles University in Prague

October 31st 2011, Prague
Outline

• Styles of annotating coordinations
  • Topological styles
  • Labeling styles

• Transformation of styles

• Evaluation
Participants of coordination

- **conjunct**
- **delimiter** (separates two conjuncts)
  - Coordinating conjunction
  - Comma or other punctuation (semicolon)
- **shared modifier** (modifies two or more conjuncts)

Examples:

- **lazy** dogs, cats and rats  more than two conjuncts (“multi-conjunct c.”)
- **Mary** came home and cried  *home* is a “private modifier”
- **John** and **Mary** or **Peter**  embedded (nested) coordinations
- **big** and **cheap** apples and **oranges**  coordinated shared modifier
Participants of coordination

- **conjunct**
- **delimiter** (separates two conjuncts)
  - Coordinating conjunction
  - Comma or other punctuation (semicolon)
- **shared modifier** (modifies two or more conjuncts)

**Examples:**
- **lazy** **dogs, cats and rats** more than two conjuncts (“multi-conjunct c.”)
- **Mary** **came home and cried** *home* is a “private modifier”
- **John and Mary or Peter** embedded (nested) coordinations
- **big and cheap** **apples and oranges** coordinated shared modifier
- Don't worry, be happy, keep smiling! As well as try hard to analyze it etc.
Topological styles (family)

Main “family” – configuration of conjuncts

Prague
- dogs
- and
- cats
- rats

Moscow
- dogs
- cats
- and

Stanford
- dogs
- cats
- and
- rats
- and
- rats
Topological styles (head)

Choice of head (which delimiter/conjunct to choose):

rightmost

leftmost
Topological styles (head)

Choice of head (which delimiter/conjunct to choose):

rightmost

leftmost

Prague  Moscow  Stanford
Topological styles (head)

Choice of head: leftmost, rightmost or mixed
Topological styles (shared modifiers)

Attachment of **shared modifiers**:

below **the head**

- lazy **dogs**,** cats**,** rats**

below **the nearest conjunct**

- **dogs**,** cats**,** rats**

- lazy
Topological styles (shared modifiers)

Attachment of shared modifiers:

below the head

below the nearest conjunct

Prague

Stanford
Topological styles (conjunction)

Attachment of coordinating conjunctions:

“between” conjuncts

below the previous conjunct

following conjunct

Stanford, head=rightmost
Topological styles (conjunction)

Attachment of coordinating conjunctions:

“between” conjunctions

below the previous conjunct

following conjunct

Moscow, head=leftmost
Topological styles (conjunction)

Attachment of coordinating conjunctions:

“between” conjuncts

“as the head” for Prague (the only applicable)

below the previous conjunct

following conjunct

Moscow, head=leftmost
Topological styles (punctuation)

Attachment of punctuation delimiters:

“between” conjuncts

below the previous conjunct

following conjunct

Prague
Topological styles – overview

How many treebanks use the given style?

• **Family** (Prague=13, Moscow=5, Stanford=6)
• **Head** (Leftmost=11, Rightmost=13, Mixed=0)
• **Shared m.** (below Head=7, Nearest conjunct=?)
• **Conjunctions** (Previous=2, Following=1, Between=8, as Head=13)
• **Punctuation** (Previous=7, Following=1, Between=14)

How many possible styles?
more than $2*3*2*3*3 + 1*3*2*1*3 = 126$ * labeling variants
Labeling styles

- Parsers can produce labeled edges (nodes): Sb, Obj, Atr,...
  - We can define special labels for coordinations (COORD, CC,...)
  - or encode additional attributes (is_member) into the dependency label: Sb_M, Obj_M, Atr_M,...
  - Some additional attributes can be deduced or guessed.
- Should we mark shared modifiers?
- Should we mark conjunct (except if head)?
- Should we mark nested coordinations (co-index modifiers)?
- In Prague styles, where to store the dependency relation of the whole coordination?
  - with conjuncts (they can have different labels in PDT)
  - with the head.
Transformations of styles

Subtasks

1) Detect coordinations in a sentence (esp. boundaries of nested coordinations)
2) Classify participants of coordinations (conjunct, commas, conjunctions, shared m.)
3) Transform each coordination to the target style
Problematic cases

Prague

Moscow
Problematic cases

“Save money, don't phone, use fax.”

PDT 2.0
Evaluation

- Transform both train and test set
  - train the parser on the transformed train set
  - evaluate on the transformed test set

- Transform only the train data
  - train the parser on the transformed train set
  - apply inverse transformation
  - evaluate on the original test set
### Preliminary results

### Questions?

Thank you.