Annotating Coordination in the Penn Treebank

Wolfgang Maier[1], Sandra Kübler[2], Erhard Hinrichs[3] and Julia Krivanek[3]

[1]University of Düsseldorf
[2]Indiana University
[3]University of Tübingen

The LAW VI at ACL 2012
Jeju, Korea
July 13, 2012

Definitions

Coordination

- A syntactic structure consisting of two or more elements (*conjuncts*), with one or more conjuncts typically, but not always preceded by a coordinating conjunction (*and*, *or*, *neither...nor*, and *but*).
- Conjuncts typically of the same category, or *unlike coordination*

Definitions

Coordination

- A syntactic structure consisting of two or more elements (*conjuncts*), with one or more conjuncts typically, but not always preceded by a coordinating conjunction (*and*, *or*, *neither...nor*, and *but*).
- Conjuncts typically of the same category, or *unlike coordination*

Conjunct

- Lexical or phrasal material of any kind
- Typically, but not necessarily, a complete constituent

Examples

(1) a. Leslie and Sandy

Examples

(1) a. Leslie and Sandy
b. Loch Ness is a lake in Scotland and famous for its monster (*unlike coordination*)

Examples

(1) a. Leslie and Sandy
b. Loch Ness is a lake in Scotland and famous for its monster (*unlike coordination*)
c. Sandy gave a record to Sue and a book to Leslie (*non-constituent coordination*)

Examples

(1) a. Leslie and Sandy
b. Loch Ness is a lake in Scotland and famous for its monster (*unlike coordination*)
c. Sandy gave a record to Sue and a book to Leslie (*non-constituent coordination*)
d. Leslie likes bagels and Sandy donuts (*gapping structure with elliptical conjunct*)

Coordination detection and scope identification

- Coordination information is important!

Coordination detection and scope identification

- Coordination information is important!
- Consider parsing:
  - Improved coordination parsing helps overall parse quality
  - Downstream NLP applications also benefit

Previous work

- Coordinations of noun compounds ("oil and gas resources") [Nakov and Hearst, 2005],
- Coordinations of symmetrical NPs [Hogan, 2007, Shimbo and Hara, 2007],
- Coordinations of the form "A CC B" where A and B are conjuncts, and CC is an overt conjunction [Kübler et al., 2009].

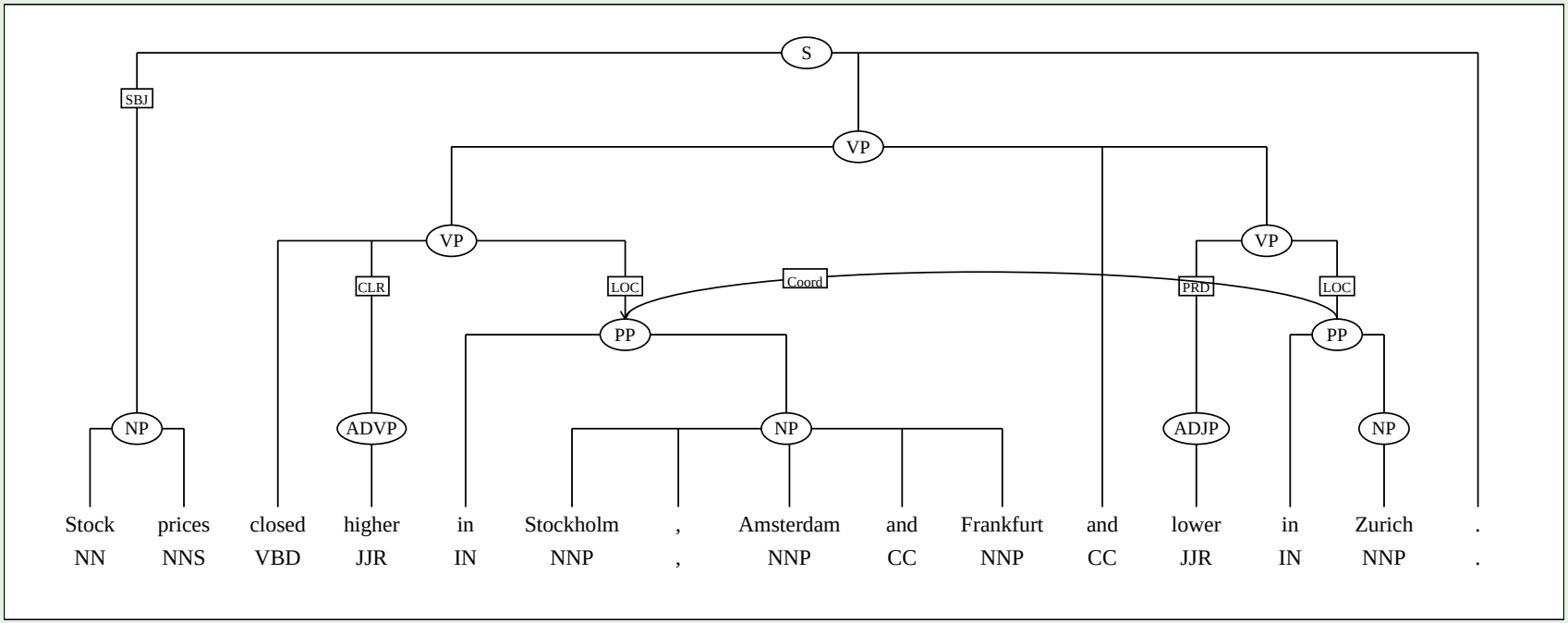
Previous work

- Coordinations of noun compounds ("oil and gas resources") [Nakov and Hearst, 2005],
- Coordinations of symmetrical NPs [Hogan, 2007, Shimbo and Hara, 2007],
- Coordinations of the form "A CC B" where A and B are conjuncts, and CC is an overt conjunction [Kübler et al., 2009].

To our knowledge, no approach for *all* coordination types

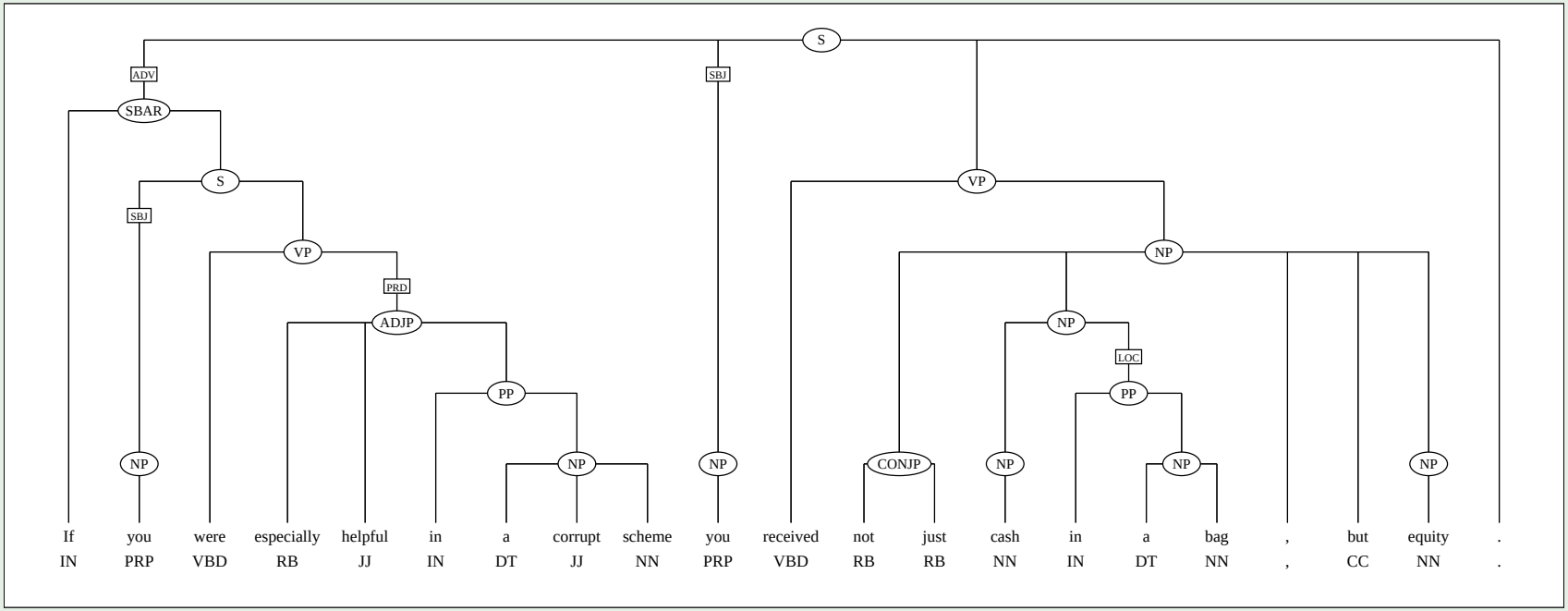
The Penn Treebank annotation principles for coordination

- Annotated on lowest level possible
- One word conjuncts coordinated on word level
- In gapped structures, symmetrical elements in conjuncts marked by gap-coindexation



The Penn Treebank annotation principles for coordination

- Multiword conjunctions grouped into `CONJP` constituents



The problem: Conjunct boundaries

- Special POS tag for coordinating conjunctions, but not for coordinating punctuation
- Handling > 2 conjuncts is difficult!

Our work

We present

Our work

We present

1. a proposal for an extended annotation of the Penn Treebank which facilitates the identification of coordination

Our work

We present

1. a proposal for an extended annotation of the Penn Treebank which facilitates the identification of coordination
2. a variety of phenomena which must be taken into account for a thorough treatment of coordination

Annotation principles

Consider a punctuation token c with left and right neighbor tokens l and r (not considering intervening coordinating conjunctions).

c is annotated as coordinating iff

Annotation principles

Consider a punctuation token c with left and right neighbor tokens l and r (not considering intervening coordinating conjunctions).

c is annotated as coordinating iff

- it is attached to the lowest common ancestor A of l and r and

Annotation principles

Consider a punctuation token c with left and right neighbor tokens l and r (not considering intervening coordinating conjunctions).

c is annotated as coordinating iff

- it is attached to the lowest common ancestor A of l and r and
- it holds that
  - the phrases directly below A which cover r, resp. l have the same label, or

Annotation principles

Consider a punctuation token c with left and right neighbor tokens l and r (not considering intervening coordinating conjunctions).

c is annotated as coordinating iff

- it is attached to the lowest common ancestor A of l and r and
- it holds that
  - the phrases directly below A which cover r, resp. l have the same label, or
  - A is labeled `UCP` (unlike coordination)

Why no automatic annotation?

Some manual decisions are necessary!

Coordination vs. apposition (1)

- Appositions can look like conjuncts
- Often have same category as modified constituent and attached at the same level

(2) The last two months have been the whole ball game , " says [$_{NP}$ Steven Norwitz] , [$_{NP}$ a vice president] . (PTB 15034)

Coordination vs. apposition (2)

Similar case: NP modification by temporal NP

(3) – : Letter from Eduard Shevardnadze to U.N. Secretary-General Perez de Cuellar , reported in [$_{NP}$ Tass] , [$_{NP-TMP}$ June 10 , 1988] . (PTB 21148)

Coordination vs. apposition (3)

- Border cases difficult to decide, especially with negated second phrase
- In case of doubt, use *and* substitution test

(4) He is [$_{NP}$ **a mechanical engineer**] , [$_{NP}$ **not an atmospheric chemist**] . (PTB 7158)

Ambiguous punctuation

- Irregular usage of commas in the PTB
- One finds "A, B, and C", "A, B and C" and even "A, and B"
- Covered by annotation guidelines

Ambiguous punctuation

(5) a. Describing itself as " asset rich , " Sea Containers said it will move immediately to sell [$_{NP}$ **two ports**] , [$_{NP}$ **various ferries**] , [$_{NP}$ **ferry services**] , [$_{NP}$ **containers**] , and [$_{NP}$ **other investments**] . (PTB 6105)

Ambiguous punctuation

(5) a. Describing itself as " asset rich , " Sea Containers said it will move immediately to sell [$_{NP}$ **two ports**] , [$_{NP}$ **various ferries**] , [$_{NP}$ **ferry services**] , [$_{NP}$ **containers**] , and [$_{NP}$ **other investments**] . (PTB 6105)

b. Stocks closed higher in [$_{NP}$ **Hong Kong**] , [$_{NP}$ **Manila**] , [$_{NP}$ **Singapore**] , [$_{NP}$ **Sydney**] and [$_{NP}$ **Wellington**] , but were lower in Seoul . (PTB 4369)

Ambiguous punctuation (2)

- Sometimes, the comma before a coordinating conjunction belongs to the preceding constituent
- Handeled by our annotation guidelines, since in the PTB, the comma is attached low to preceding constituent (e.g., an apposition)

Ambiguous punctuation (2)

(6) a. Berthold [$_{VP}$ is based in Wildbad , West Germany ,] and [$_{VP}$ also has operations in Belgium] . (PTB 4988)

Ambiguous punctuation (2)

(6) a. Berthold [$_{VP}$ is based in Wildbad , West Germany ,] and [$_{VP}$ also has operations in Belgium] . (PTB 4988)

b. . . . Gillette South Africa will sell [$_{NP}$ manufacturing facilities in Springs , South Africa ,] and [$_{NP}$ its business in toiletries and plastic bags] to Twins Pharmaceuticals Ltd. . . . (PTB 6154)

Ambiguous punctuation (2)

(6) a. Berthold [$_{VP}$ is based in Wildbad , West Germany ,] and [$_{VP}$ also has operations in Belgium] . (PTB 4988)

b. . . . Gillette South Africa will sell [$_{NP}$ manufacturing facilities in Springs , South Africa ,] and [$_{NP}$ its business in toiletries and plastic bags] to Twins Pharmaceuticals Ltd. . . . (PTB 6154)

c. [$_{S}$ I want white America to talk about it , too ,] but [$_{S}$ I 'm convinced that the grapevine is what 's happening] . " (PTB 10130)

Ambiguous punctuation (3)

- Also ambiguous: Coordinate structures on clausal level
- Often only uses commas or semicolons (no overt conjunctions)
- Difficult to distinguish automatically from other types of parataxis
- Again in case of doubt, use *and* substitution test

Ambiguous punctuation (3)

Annotated as coordination

(7) a. [$_S$ **In 1980 , 18 % of federal prosecutions concluded at trial**] ; [$_S$ **in 1987 , only 9 % did**] . (PTB 12113)

Ambiguous punctuation (3)

Annotated as coordination

(7) a. [$_S$ **In 1980 , 18 % of federal prosecutions concluded at trial**] ; [$_S$ **in 1987 , only 9 % did**] . (PTB 12113)

b. [$_S$ **Various ministries decided the products businessmen could produce and how much**] ; and [$_S$ **government-owned banks controlled the financing of projects and monitored whether companies came through on promised plans**] . (PTB 12355)

Ambiguous punctuation (3)

Not annotated as coordination

(8) a. [$_S$ This does n't necessarily mean larger firms have an advantage] ; [$_S$ Mr. Pearce said GM works with a number of smaller firms it regards highly] . (PTB 12108)

Ambiguous punctuation (3)

Not annotated as coordination

(8) a. [$_S$ This does n't necessarily mean larger firms have an advantage] ; [$_S$ Mr. Pearce said GM works with a number of smaller firms it regards highly] . (PTB 12108)

b. [$_S$ Senator Sasser of Tennessee is chairman of the Appropriations subcommittee on military construction] ; [$_S$ Mr. Bush 's $ 87 million request for Tennessee increased to $ 109 million] . (PTB 12223)

Non coordinative use of punctuation

- Some sentences show coordinating conjunctions in appositions.
- Example: Syntactic annotation/annotation guidelines make it look like coordination

(9) a. The NASD , which operates the Nasdaq computer system on which 5,200 OTC issues trade , compiles short interest data in [$_{NP}$ two categories] **:** [$_{NP}$ the approximately two-thirds . . . ; and the one-third . . .] . (PTB 21080)

Non-coordinative use of punctuation

- Example: Syntactic annotation/annotation guidelines identify it as a non-coordination
- Reason: Coordinating conjunction grouped under parenthetical node (`PRN`) or under fragment (`FRAG`).

(10) a. Martha was [$_{ADJP}$ pleased , [$_{PRN}$ but nowhere near as much as Mr. Engelken]] . (PTB 14598)

Non-coordinative use of punctuation

- Example: Syntactic annotation/annotation guidelines identify it as a non-coordination
- Reason: Coordinating conjunction grouped under parenthetical node (`PRN`) or under fragment (`FRAG`).

(10) a. Martha was [$_{ADJP}$ pleased , [$_{PRN}$ but nowhere near as much as Mr. Engelken]] . (PTB 14598)
b. The HUD scandals will simply [$_{VP}$ continue , [$_{FRAG}$ but under new mismanagement]] . (PTB 15629)

Coordination in NP premodification

- PTB annotation guidelines: Conjuncts project to phrase level
- Exception: NP premodifiers only project if longer than one word
- We treat all NP premodifiers as coordinating

Coordination in NP premodification

(11) a. Yesterday , it received a [$_{ADJP}$ **$ 15 million**] , [$_{JJ}$ **three-year**] contract from Drexel Burnham Lambert . (PTB 6485)

Coordination in NP premodification

(11) a. Yesterday , it received a [$_{ADJP}$ **$ 15 million**] , [$_{JJ}$ **three-year**] contract from Drexel Burnham Lambert . (PTB 6485)

b. There 's nothing in the least contradictory in all this , and it would be nice to think that Washington could tolerate a [$_{ADJP}$ **reasonably sophisticated**] , [$_{JJ}$ **complex**] view . (PTB 8018)

Coordination in NP premodification

(11) a. Yesterday , it received a [$_{ADJP}$ **$ 15 million**] , [$_{JJ}$ **three-year**] contract from Drexel Burnham Lambert . (PTB 6485)

b. There 's nothing in the least contradictory in all this , and it would be nice to think that Washington could tolerate a [$_{ADJP}$ **reasonably sophisticated**] , [$_{JJ}$ **complex**] view . (PTB 8018)

c. Perhaps the shock would have been less if they 'd fixed to another [$_{NN}$ **low-tax**] , [$_{VBN}$ **deregulated**] , [$_{JJ}$ **supply-side**] economy . (PTB 10463)

Data set

- A look at coordination statistics in the PTB
- Collected from 23,678 sentences (605,064 words), average sentence length 25.6 words

	full		av. per sentence	
	total	coord.	total	coord.
,	28,853	3,924	1.22	0.17
;	684	547	0.03	0.02
CCs	14 267		0.60	

Annotation properties

- Statistics over noun phrases
- Without the annotation, a significant number of conjuncts is lost

Number of conj.	w/ annot.	w/o annot.
2	12 689	13 917
3	2 243	1 195
4	653	220
5	234	35
6	90	18
≥ 7	94	0

Summary

Summary

1. We have treated a wide variety of coordination phenomena in English

Summary

1. We have treated a wide variety of coordination phenomena in English
2. We have proposed an extension for the Penn Treebank annotation from which NLP applications will profit

Summary

1. We have treated a wide variety of coordination phenomena in English
2. We have proposed an extension for the Penn Treebank annotation from which NLP applications will profit
3. We plan on making the annotation publicly available

Thank you for your attention!

Questions?

References I

Hogan, D. (2007).

Coordinate noun phrase disambiguation in a generative parsing model.
In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 680–687, Prague, Czech Republic.

Kübler, S., Hinrichs, E. W., Maier, W., and Klett, E. (2009).

Parsing coordinations.
In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 406–414, Athens, Greece.

Nakov, P. and Hearst, M. (2005).

Using the web as an implicit training set: Application to structural ambiguity resolution.
In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 835–842, Vancouver, Canada.

Shimbo, M. and Hara, K. (2007).

A discriminative learning model for coordinate conjunctions.
In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 610–619, Prague, Czech Republic.