Annotating Coordination in the Penn Treebank

Wolfgang Maier
Universität Düsseldorf
Institut für Sprache und Information
maierw@hhu.de

Sandra Kübler
Indiana University
Department of Linguistics
skuebler@indiana.edu

Erhard Hinrichs
Universität Tübingen
Seminar für Sprachwissenschaft
eh@sfs.uni-tuebingen.de

Julia Krivanek
Universität Tübingen
Seminar für Sprachwissenschaft
julia.krivanek@student.uni-tuebingen.de

Abstract

Finding coordinations provides useful information for many NLP endeavors. However, the task has not received much attention in the literature. A major reason for that is that the annotation of major treebanks does not reliably annotate coordination. This makes it virtually impossible to detect coordinations in which two conjuncts are separated by punctuation rather than by a coordinating conjunction. In this paper, we present an annotation scheme for the Penn Treebank which introduces a distinction between coordinating from non-coordinating punctuation. We discuss the general annotation guidelines as well as problematic cases. Eventually, we show that this additional annotation allows the retrieval of a considerable number of coordinate structures beyond the ones having a coordinating conjunction.

1 Introduction

1.1 Motivation

Coordination is a difficult topic, in terms of linguistic description and analysis as well as for NLP approaches. Most linguistic frameworks still struggle with finding an account for coordination that is descriptively fully adequate (Hartmann, 2000). This is also the reason why coordination is not adequately encoded in the annotation of major treebanks. From an NLP perspective, coordination is one of the major sources for errors in parsing (Hogan, 2007). If parsing of coordinate structures can be improved, overall parsing quality also benefits (Kübler et al., 2009). And consequently, downstream NLP applications, such as question answering or machine translation, would benefit as well.

However, since linguistic frameworks in general are challenged by the diverse phenomena of coordination, a consistent annotation of coordinate structures, clearly marking the phenomenon as such as well as its scope, is a difficult enterprise. Consequently, this makes the detection of conjuncts and their boundaries a highly non-trivial task. Nevertheless, an exact detection of coordination scopes is necessary for improving parsing approaches to this phenomenon.

A first step in the detection of the single conjuncts of a coordinate structure is a reliable detection of the presence of a coordinate structure as such and of the boundaries between its conjuncts. One highly predictive marker for the detection of coordinate structures is the presence of a coordinating conjunction such as `and`, `or`, `neither...nor`, and `but`. In treebanks, coordinating conjunctions are generally easy to identify by a specialized part of speech (POS) tag, for instance `CC` in the Penn Treebank (PTB) (Marcus et al., 1993) and `KON` in the Stuttgart-Tübingen tagset (STTS) (Thielen and Schiller, 1994). However, if the coordinate structure has more than 2 conjuncts, or if it is on the clause level, the conjuncts are separated by punctuation signs such as commas rather than by overt coordinating conjunctions. In the PTB, they are annotated with the POS tag `,`; in the German treebanks, TIGER (Brants et al., 2002), Negra (Skut et al., 1998), TüBa-D/S (Hinrichs et al., 2000), and TüBa-D/Z (Telljohann et al., 2004) using the STTS,

Proceedings of the 6th Linguistic Annotation Workshop, pages 166–174,
Jeju, Republic of Korea, 12-13 July 2012.

they are annotated with the POS tags `$,` and `$:`, like all other punctuation without coordinating function.

Automatically identifying coordinate structures and the scope of their conjuncts in the Penn Treebank is challenging since coordinate structures as a whole and their conjuncts are not explicitly marked in the annotation by special phrasal or lexical nodes. Figure 1 shows an example sentence with two coordinate structures, the inside one a coordinate noun phrase (NP) with 3 conjuncts, and the outside one a coordinate verb phrase (VP) with two complex conjuncts. These coordinate structures are labeled by ordinary phrasal categories such as VP and NP and can thus not be distinguished at the phrasal level from VPs and NP that do not involve coordination.

There are approaches to improving parsing for coordinations, but most of these approaches are restricted to very narrow definitions such as coordinations of noun compounds such as "oil and gas resources" (Nakov and Hearst, 2005), coordinations of symmetrical NPs (Hogan, 2007; Shimbo and Hara, 2007), or coordinations of "A CC B" where A and B are conjuncts, and CC is an overt conjunction (Kübler et al., 2009). To our knowledge, there is no attempt at covering all coordination types.

One goal of this paper is to demonstrate a wide range of coordination phenomena that have to be taken into account in a thorough treatment of coordinations. We additionally present a proposal for an enhanced annotation of coordination for the Penn Treebank. The annotation is focused on punctuation and allows for an in-depth investigation of coordinations, for example for linguistic treatments, but also for work on coordination detection, from which many NLP applications can profit.

The structure of this paper is as follows. In section 2, we look at syntactic treatments of coordination, and we have a look at the Penn Treebank guidelines. Section 3 is dedicated to the presentation of a "stylebook" for the enhanced annotation of coordination that we advocate in the present paper. We outline our annotation decisions and the issues that we encountered. Section 4 contains an empirical analysis of the coordinations in the PTB, made possible by the new annotation. Finally, section 5 concludes the paper.

2 Related Work

2.1 Coordination in Linguistics

Coordinations are complex syntactic structures that consist of two or more elements (conjuncts), with one or more conjuncts typically, but not always preceded by a coordinating conjunction such as `and`, `or`, `neither...nor`, and `but`. However, see section 3 for examples of coordinations that lack coordinating conjunctions altogether. Coordinate structures can conjoin lexical and phrasal material of any kind and typically exhibit syntactic parallelism in the sense that each conjunct belongs to the same lexical or phrasal category. However, coordinations of unlike categories such as `Loch Ness is a lake in Scotland and famous for its monster` are also possible. The conjuncts are typically syntactic constituents; in fact, coordinate structures are among the classic constructions used to test for constituency. However, there are well-known cases of non-constituent conjunctions such as `Sandy gave a record to Sue and a book to Leslie` and gapping structures with one or more elliptical conjuncts such as `Leslie likes bagels and Sandy donuts`. Incidentally, the coordinate structure in Figure 1 consitutes an example of non-constituent conjunction since the second conjunct `lower in Zurich` does not form a single constituent. The PTB treats this conjunct as a VP. However, note that the conjunct is not headed by a verb; rather the verb is elided.

It is precisely the wide range of distinct subcases of constituent structures that makes their linguistic analysis challenging and that makes it hard to construct adequate language models for the computational processing of coordinate structures. The purpose of the present paper is not to refine existing theoretical accounts of coordinate structures such as those proposed in Generative Grammar, Generalized Phrase Structure Grammar, Head-Driven Phrase Structure Grammar, Tree Adjoining Grammar, or Dependency Grammar. Rather, our goal is a much more modest one and focuses on written language only, where punctuation is among the reliable cues for predicting cases of coordinate structures and for identifying the boundaries of individual conjuncts, especially for coordinate structures with

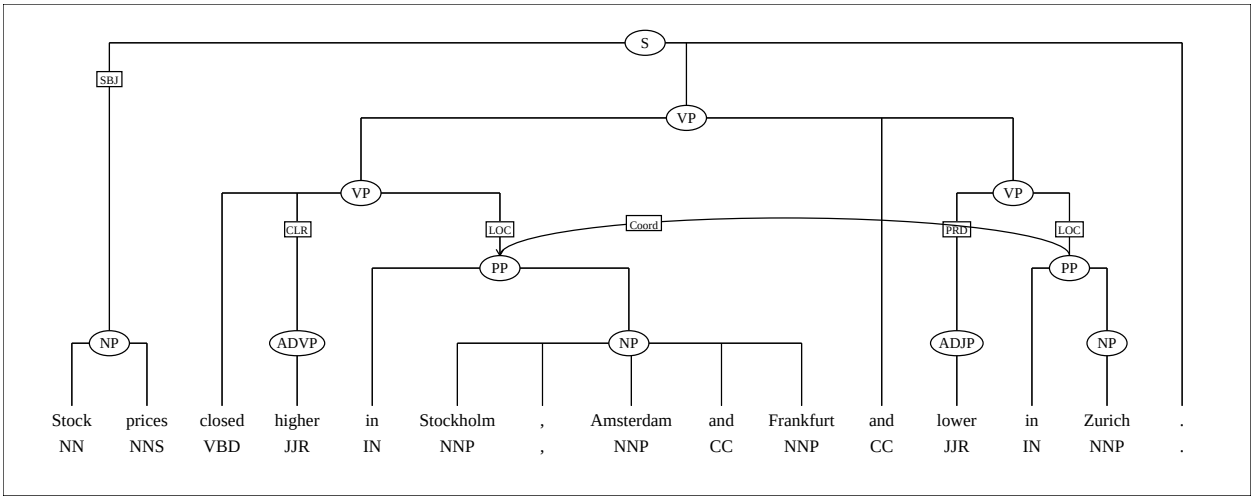

Figure 1: An example with two embedded coordinations.

more than two conjuncts, which have been largely ignored in computational modeling of language thus far.

Since supervised models for statistical parsing require annotated training material, we will propose a more fine-grained annotation scheme for punctuation than has thus far been incorporated into existing treebanks. The present paper focuses on English only and will use the Penn Treebank Bracketing Guidelines as the annotation scheme for which such more fine-grained annotations will be proposed. However, the proposed modifications can be easily imported to other treebanks for English such as CCGBank or treebanks for other language, and we conjecture that they would lead to improved language models for coordinate structures for those treebanks as well.

In order to properly ground the discussion, we will now review Penn Treebank Bracketing Guidelines.

2.2 Penn Treebank Guidelines

The Penn Treebank Bracketing Guidelines (Bies et al., 1995, sec. 7) describe extensively how to treat coordination in terms of bracketing. The guidelines state that coordinate structures are annotated on the lowest level possible. One word conjuncts are coordinated on the word level. An example for this is shown in Figure 1 in the coordinated NP `Stockholm , Amsterdam and Frankfurt`. In gapped structures, symmetrical elements in the conjuncts are marked using gap-coindexation. In the example in Figure 1, the coindexation is shown as a secondary edge from the prepositional phrase (PP) in the second conjunct to the PP in the first one.

The guidelines also discuss multi-word coordinating conjunctions such as `as well as` or `instead of` and discontinuous conjunctions such as `not only ...but` or `not ...but instead`. Multi-word coordinating conjunctions, including discontinuous ones, are grouped into `CONJP` constituents. Single word portions of discontinuous conjunctions are not marked as such. Figure 2 shows an example of a discontinous coordinating conjunction in which the first part is projected to a `CONJP` while the second part is a single word and thus not projected.

The manual does not mention coordinate structures with more than 2 conjuncts or without overt conjunctions, and the only examples in which the comma takes over the role of a coordinating conjunction refer to "difficult cases" such as the sentence in Figure 3, in which symmetry is enforced by anti-placeholders `*NOT*`.

3 Annotation of Coordinating Punctuation

We annotate all intra-sentential punctuation in the Penn Treebank and determine for each punctuation sign whether it is part of a coordination or not. As far as possible, decisions are based on the syntactic

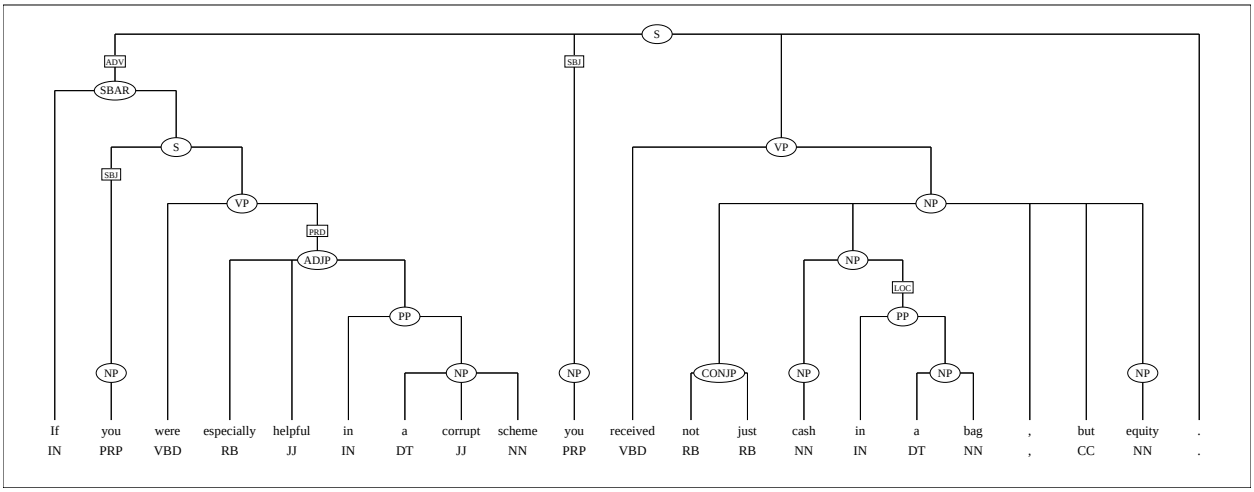

Figure 2: An example with a multi-word conjunction.

annotations in the treebank.

3.1 Annotation principles

The principal guidelines for the enhanced annotation of coordination are as follows. Let t be a punctuation token and let t_l and t_r be the tokens immediately on the left and the right of t (disregarding coordinating conjunctions). We annotate t as coordinating iff

1. t is attached to the lowest node t_c which dominates both t_l and t_r, and

2a. in the symmetrical case: the non-terminals directly dominated by t_c which also dominate t_l, resp. t_r, have the same label;

2b. in the asymmetrical case: t_c is labeled `UCP` (coordination of unlike constituents) – or t_c is `S`, and the two non-terminals dominating t_l and t_r are different (since coordination of unlike clausal constituents is grouped under an `S` rather than a `UCP`).

In cases where there are no nodes between t and t_c, we check the POS tags of t_l and t_R for equality. In theory, these two rules, given the syntactic annotation, should be sufficient to find all cases of coordination. However, in practice, the situation is more complicated, as shown in the next subsection.

For example, in Figure 1, the comma is labeled as coordination since the two words to the left and right are directly dominated by an NP, and they both have the same POS tag, NNP, and thus follow rule 2a. The comma in Figure 2 is also annotated as a coordination following 2a since the words to the left and right are both dominated by NPs, as is the node dominating all words in question. We present examples for symmetrical coordinations on the clausal and phrasal level in (1).

(1)
a. Richard Stoltzman has taken a $[_{JJR}$ gentler$]$, $[_{ADJP}$ more audience-friendly approach$]$. (PTB 3968)
b. The two leaders are expected to discuss $[_{NP}$ changes sweeping the East bloc$]$ as well as $[_{NP}[_{NP}$ human-rights issues$]$, $[_{NP}$ regional disputes$]$ and $[_{NP}$ economic cooperation$]]$. (PTB 6798)
c. These critics are backed by several academic studies showing that the adoption of poison pills reduces shareholder values not merely $[_{PP}$ in the short run$]$, but also $[_{PP}$ over longer periods$]$. (PTB 5056)
d. Our pilot simply $[_{VP}$ laughed$]$, $[_{VP}$ fired up the burner$]$ and $[_{VP}$ with another blast of flame lifted us , oh , a good 12-inches above the water level$]$. (PTB 4465)
e. $[_{S}$ He believes in what he plays$]$, and $[_{S}$ he plays superbly$]$. (PTB 3973)
f. $[_{S}$ Dow Jones industrials 2596.72 , off 17.01$]$; $[_{S}$ transportation 1190.43 , off 14.76$]$; $[_{S}$ utilities 215.86 , up 0.19$]$.

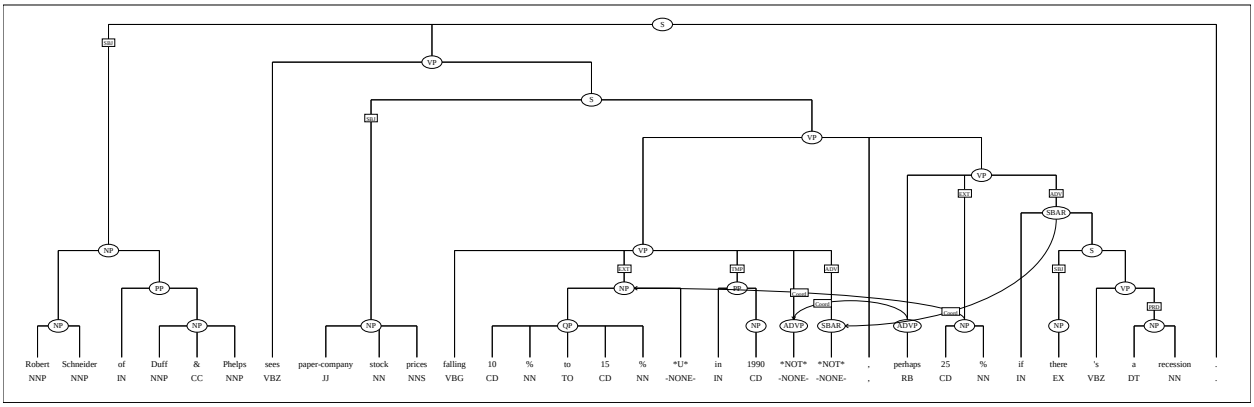

Figure 3: A difficult coordination example.

(PTB 13082)

The examples in (2) show cases of coordination of unlike constituents. These cases are covered by the rule 2b described above; in the first two sentences, all conjuncts are dominated by `UCP`, and the last sentence is an example of a clausal coordination, that is projected to an `S` node.

(2) a. Assuming final enactment this month , the prohibition will take effect [$_{ADVP}$ 96 days later] , or [$_{PP}$ in early February] . (PTB 6499)

b. My wife and I will stay [$_{PP}$ through the skiing season] , or [$_{SBAR}$ until the money runs out] – whichever comes first . (PTB 15255)

c. This perhaps was perceived as [$_{NP}$ a " bold " stance] , and thus [$_{ADJP}$ suspicious] . (PTB 18051)

d. [$_{S}$ Mr. Trotter 's painting showed a wall of wood boards with painted ribbons tacked down in a rectangle] ; [$_{SINV}$ tucked behind the ribbons were envelopes , folded , faded and crumpled papers and currency] . (PTB 8698)

The example in (3) shows a comma that has two different functions: The comma before `and` delimits the relative clause modifying `oral orders`, and at the same time marks the coordination. Since we are interested in all cases of coordination, such multi-functional punctuation marks are annotated as coordinations if that is one of their functions.

(3) The affected practices include [$_{NP}$ the placing of oral orders , which is the way most public customer orders are placed] , and [$_{NP}$ trading between affiliated brokers] , even though in some cases trading with affiliates may be the only way to obtain the best execution for a client . (PTB 15541)

3.2 Problematic Cases

Coordination vs. apposition In many cases, appositions show the same characteristics as the rules above. An apposition is not restricted to be of the same category as the constituent it modifies, but in many cases, it is. These cases are the main reason for the manual annotation since they cannot be distinguished automatically. Thus, if the second phrase defines or modifies the first one, we do not annotate the intervening commas as coordination. An example for an apposition that follows the rules above is given in (4).

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

The same holds for cases in which a temporal NP modifies another NP, such as in example (5). Here, the NP `Tass` is modified by the temporal NP `June 10 , 1988`.

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

There are cases, especially ones in which the second phrase is negated, for which it is difficult to decide between coordination and apposition. The sentence in (6) shows an example. For these cases, we decided to treat them as coordination.

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

Ambiguous punctuation Commas before coordinating conjunctions are typically signs of coordination. Note that the usage of commas in the Penn Treebank is not very regular, and cases of "A, B, and C" can be found along with cases of "A, B and C" and cases of "A, and B", as shown in the examples in (7). All these cases are covered by rule 2a.

(7) 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)

c. [$_{NP}$ Sidley & Austin , a leading Chicago-based law firm] , and [$_{NP}$ Ashurst Morris Crisp , a midsized London firm of solicitors] , are scheduled today to announce plans to open a joint office in Tokyo . (PTB 5367)

However, there are also cases in which the comma before a coordinating conjunction is clearly not part of the coordination, but rather belongs to the preceding constituent, such as in the examples in (8). In these cases, the syntactic annotation shows that the comma is not a coordination comma by attaching it low to the preceding constituent; we do not annotate these commas as coordination phenomena.

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

b. Under the plan , 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. , an affiliate of Anglo American Corp. , a South African company . (PTB 6154)

c. [$_{S}$ Last week 's uncertainty in the stock market and a weaker dollar triggered a flight to safety] [$_{PRN}$, he said ,] [$_{S}$ but yesterday the market lacked such stimuli] . (PTB 8252)

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

Another ambiguous case can be found in coordinate structures on the clausal level, which often does not use overt coordinating conjunctions, but rather commas or semicolons. These cases of coordination are difficult to distinguish automatically from other types of parataxis. The examples in (9) we regard as coordinations while the examples in (10) are not since the relation between them is elaborative.

(9) 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)

(10) 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 conjunctions There are sentences that involve coordinating conjunctions in structures that are not coordinations but rather ap-

positions. While the first example in (11) cannot be distinguished from coordination based on our annotation guidelines (cf. sec. 3.1) and the syntactic annotation, the syntactic annotation for the other two sentences shows that these are not considered cases of coordination, either by grouping the coordinating conjunction under a parenthetical node (PRN) or under a fragment (FRAG).

(11) a. The NASD , which operates the Nasdaq computer system on which 5,200 OTC issues trade , compiles short interest data in [$_{NP}$ [$_{NP}$ two categories] : [$_{NP}$ the approximately two-thirds , and generally biggest , Nasdaq stocks that trade on the National Market System ; and the one-third , and generally smaller , Nasdaq stocks that are n't a part of the system]] . (PTB 21080)

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

c. The HUD scandals will simply [$_{VP}$ continue , [$_{FRAG}$ but under new mismanagement]] . (PTB 15629)

Coordination in NP premodification The Penn Treebank Bracketing Guidelines (Marcus et al., 1993) state that generally conjuncts are projected to the phrase level before they are coordinated. There is one exception: premodifiers in NPs, which are only projected if they consist of more than one word. In such cases, it is not obvious from the tree that there is a coordination. But even if there is no explicit marking of coordination in the syntactic analysis, we do annotate the coordination. Examples are shown in (12).

(12) 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)

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

Table 1: Annotation of punctuation

4 Properties of the Annotation

For the empirical analysis presented here, we use approximately half the Penn Treebank. The data set has a size of 23 678 sentences and 605 064 words in total, with an average length of 25.6 words per sentence.

Table 1 shows some basic statistics, more specfically:

1. the numbers of annotated commas, semicolons, and coordinating conjunctions (CC) and their total numbers over the entire data set, and

2. the average numbers of annotated commas, semicolons, and coordinating conjunctions (CC) and their average number per sentence.

The numbers show that approximately 14% of all commas and 80% of all semicolons are used in coordinate structures. CCs constitute only 2.36% of all words. If we count CCs as well as the punctuation signs that are annotated as being part of a coordination, the number rises to 3.10% of all words. These numbers show that we cannot assume that all sentence-internal punctuation is related to coordination, but that the use of commas and semicolons to separate conjuncts is not a marginal phenomenon.

Table 2 offers a first look at the distribution of the number of conjuncts that occur in coordinate structures. Our present investigation focuses exclusively on noun phrase coordination. Given that, in principle, our annotation marks all conjunctions, and given that our annotation guidelines state that all conjuncts must be sisters, it is rather straightforward to determine the number of conjuncts of a coordination: We simply count the separators between conjuncts, i.e. CCs and conjunction punctuation, below a given non-terminal while counting

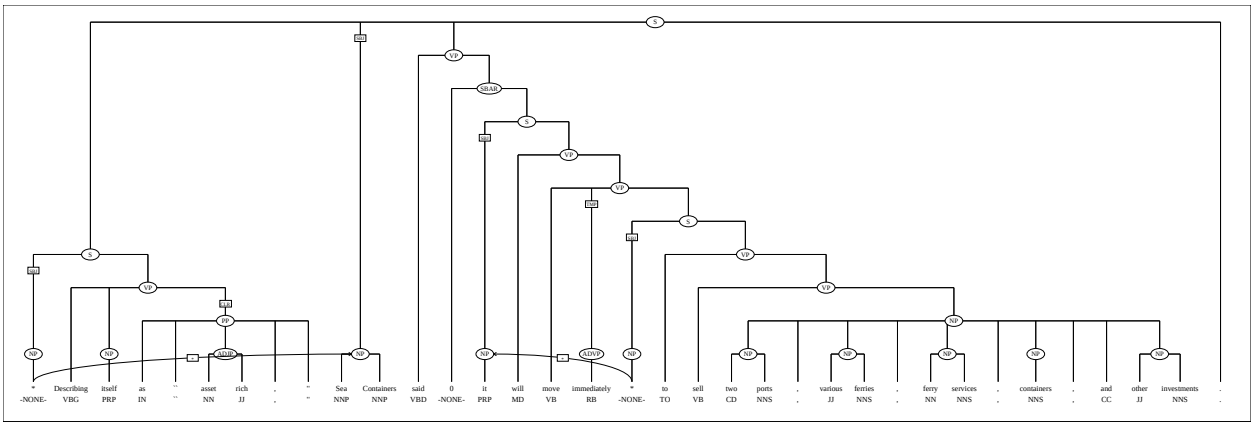

Figure 4: An example with more than two conjuncts.

No. 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

Table 2: Number of conjuncts below NX/NP

adjacent separators as singles (in order to count sequences of "A, and B" as single separator). The number of conjuncts is then the number of separators plus one. Without annotation, i.e. when only considering CC, we find that 2 882 sentences (12% out of the total of 23 678 sentences) have coordinations consisting of more than two conjuncts. If we additionally consider the coordination punctuation, this number rises significantly to 4 764 (20%). When looking at noun phrase coordination, more precisely, at coordination below NX and NP, the added value of our enhanced coordination annotation is especially apparent: It is clear from the numbers in Table 2 that we would miss a high number of coordinations, especially multi-conjunct structures, without the annotation and that this additional number of coordinations can be found reliably using our enhanced annotation.

As an example for a coordination that would be difficult to identify correctly, consider sentence (7-a). The syntactic annotation is shown in Fig. 4. While the CC tag on the last and would allow for the identification of the coordination of containers, and other investments, all NPs which are in front of those two could not be recognized as part of the coordination. A more detailed investigation of coordinate structures beyond noun phrases that would also include an assessment of the scope of coordinations is left for future work.

5 Conclusion

In this paper, we have listed a wide range of coordination phenomena that have to be taken into consideration for an exhaustive treatment of coordination. We have presented a new annotation layer for the Penn Treebank which allows for a thorough treatment of coordination by clarifying the status of punctuation. Furthermore, in an empirical analysis, we have demonstrated the utility of our annotation, showing that it allows for the detection of a large number of coordinations which cannot be detected when only coordinating conjunctions are considered.

The annotation opens the way for interesting future research. We are pursuing two different paths. On the one hand, we are investigating possibilites for the identification of coordination scope, also beyond phrasal coordination. For the first time, this is now possible using supervised methods. On the other hand, we are working on improving parser performance on coordinations on the basis of our new annotation, beyond the restricted types of coordination considered in previous works.

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre, 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. University of Pennsylvania.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In Erhard Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgaria.

Katharina Hartmann. 2000. *Right Node Raising and Gapping*. John Benjamins, Amsterdam, The Netherlands.

Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. 2000. The Verbmobil treebanks. In *Proceedings of KONVENS 2000, 5. Konferenz zur Verarbeitung natürlicher Sprache*, pages 107–112, Ilmenau, Germany.

Deirdre Hogan. 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.

Sandra Kübler, Erhard W. Hinrichs, Wolfgang Maier, and Eva Klett. 2009. Parsing coordinations. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 406–414, Athens, Greece.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. Special Issue on Using Large Corpora: II.

Preslav Nakov and Marti Hearst. 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.

Masashi Shimbo and Kazuo Hara. 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.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper texts. In *ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2229–2235, Lisbon, Portugal.

Christine Thielen and Anne Schiller. 1994. Ein kleines und erweitertes Tagset fürs Deutsche. In Helmut Feldweg and Erhard Hinrichs, editors, *Lexikon & Text*, pages 215–226. Niemeyer, Tübingen.