Coordination Annotation Extension in the Penn Tree Bank

Jessica Ficler
Computer Science Department
Bar-Ilan University
Israel
jessica.ficler@gmail.com

Yoav Goldberg
Computer Science Department
Bar-Ilan University
Israel
yoav.goldberg@gmail.com

Abstract

Coordination is an important and common syntactic construction which is not handled well by state of the art parsers. Coordinations in the Penn Treebank are missing internal structure in many cases, do not include explicit marking of the conjuncts and contain various errors and inconsistencies. In this work, we initiated manual annotation process for solving these issues. We identify the different elements in a coordination phrase and label each element with its function. We add phrase boundaries when these are missing, unify inconsistencies, and fix errors. The outcome is an extension of the PTB that includes consistent and detailed structures for coordinations. We make the coordination annotation publicly available, in hope that they will facilitate further research into coordination disambiguation. [1]

1 Introduction

The Penn Treebank (PTB) (Marcus et al., 1993) is perhaps the most commonly used resource for training and evaluating syntax-based natural language processing systems. Despite its widespread adoption and undisputed usefulness, some of the annotations in PTB are not optimal, and could be improved. The work of Vadas and Curran (2007) identified and addressed one such annotation deficiency – the lack of internal structure in base NPs. In this work we focus on the annotation of coordinating conjunctions.

Coordinating conjunctions (e.g. *"John **and** Mary"*, *"to be **or** not to be"*) are a very common syntactic construction, appearing in 38.8% of the sentences in the PTB. As noted by Hogan (2007), coordination annotation in the PTB are not consistent, include errors, and lack internal structure in many cases (Hara et al., 2009; Hogan, 2007; Shimbo and Hara, 2007). Another issue is that PTB does not mark whether a punctuation is part of the coordination or not. This was resolved by Maier et al. (2012) which annotated punctuation in the PTB .

These errors, inconsistencies, and in particular the lack of internal structural annotation turned researchers that were interested specifically in coordination disambiguation away from the PTB and towards much smaller, domain specific efforts such as the Genia Treebank (Kim et al., 2003) of biomedical texts (Hara et al., 2009; Shimbo and Hara, 2007).

In addition, we also find that the PTB annotation make it hard, and often impossible, to correctly identify the elements that are being coordinated, and tell them apart from other elements that may appear in a coordination construction. While most of the coordination phrases are simple and include only conjuncts and a coordinator, many cases include additional elements with other syntactic functions , such as markers (e.g. *"**Both** Alice and Bob"*), connectives (e.g. *"Fast and **thus** useful"*) and shared elements (e.g. *"**Bob's** principles and opinions"*) (Huddleston et al., 2002). The PTB annotations do not differentiate between these elements. For example, consider the following coordination phrases which begin with a PP:

(a) *"[in the open market]$_{PP}$, [in private transactions] or [otherwise]."*
(b) *"[According to Fred Demler]$_{PP}$, [Highland Valley has already started operating] and [Cananea is expected to do so soon]."*

Even though the first element is a conjunct only in (a), both phrases are represented with the

[1] The data is available in:
https://github.com/Jess1ca/CoordinationExtPTB

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 834–842,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

marked elements as siblings.

Our goal in this work is to fix these deficiencies. We aim for an annotation in which:

- All coordination phrases are explicitly marked and are differentiated from non-coordination structures.
- Each element in the coordination structure is explicitly marked with its role within the coordination structure.
- Similar structures are assigned a consistent annotation.

We also aim to fix existing errors involving coordination, so that the resulting corpus includes as few errors as possible. On top of these objectives, we also like to stay as close as possible to the original PTB structures.

We identify the different elements that can participate in a coordination phrase, and enrich the PTB by labeling each element with its function. We add phrase boundaries when these are missing, unify inconsistencies, and fix errors. This is done based on a combination of automatic processing and manual annotation. The result is an extension of the PTB trees that include consistent and more detailed coordination structures. We release our annotation as a diff over the PTB.

The extended coordination annotation fills an important gap in wide-scale syntactic annotation of English syntax, and is a necessary first step towards research on improving coordination disambiguation.

2 Background

Coordination is a very common syntactic structure in which two or more elements are linked. An example for a coordination structure is *“Alice and Bob traveled to Mars”*. The elements (*Alice* and *Bob*) are called the *conjuncts* and *and* is called the *coordinator*. Other coordinator words include *or*, *nor* and *but*. Any grammatical function can be coordinated. For examples: *“[relatively active]$_{ADJP}$ but [unfocused]$_{ADJP}$”* ; *“[in]$_{IN}$ and [out]$_{IN}$ the market”*. While it is common for the conjuncts to be of the same syntactic category, coordination of elements with different syntactic categories are also possible (e.g. *“Alice will visit Earth [tomorrow]$_{NP}$ or [in the next decade]$_{PP}$”*).

Less common coordinations are those with non-constituent elements. These are cases such as *“equal to or higher than”*, and coordinations from the type of Argument-Cluster (e.g. *“Alice has visited 4 planets in 2014 and 3 more since then”*) and Gapping (e.g. *“Bob lives in Earth and Alice in Saturn”*) (Dowty, 1988).

2.1 Elements of Coordination Structure

While the canonical coordination cases involve conjuncts linked with a coordinator, other elements may also take part in the coordination structure: markers, connective adjectives, parentheticals, and shared arguments and modifiers. These elements are often part of the same syntactic phrase as the conjuncts, and should be taken into account in coordination structure annotation. We elaborate on the possible elements in a coordination phrase:

Shared modifiers Modifiers that are related to each of the conjuncts in the phrase. For instance, in *“Venus’s density and mean temperature are very high”*, *Venus’s* is a shared modifier of the conjuncts *“density”* and *“mean temperature”* [2].

Shared arguments Phrases that function as arguments for each of the conjuncts. For instance, in *“Bob cleaned and refueled the spaceship.”*, *“the spaceship”* and *“Bob”* are arguments of the conjuncts *cleaned* and *refuel* [3].

Markers Determiners such as *both* and *either* that may appear at the beginning of the coordination phrase (Huddleston et al., 2002). As for example in *“Both Alice and Bob are Aliens”* and *“Either Alice or Bob will drive the spaceship”*. In addition to the cases documented by Huddleston et al, our annotation of the Penn Treebank data reveals additional markers. For examples: *“between 15 million and 20 million* ; *“first and second respectively*”.

Connective adjectives Adverbs such as *so*, *yet*, *however*, *then*, etc. that commonly appear right after the coordinator (Huddleston et al., 2002). For instance *“We plan to meet in the middle of the way and then continue together”*.

Parenthetical Parenthetical remarks that may appear between the conjuncts. For examples:

[2] Here, the NP containing the coordination (“Venus’s density and mean temperature”) is itself an argument of “are very high”.

[3] While both are shared arguments, standard syntactic analyses consider the subject (Bob) to be outside the VP containing the coordination, and the direct object (the spaceship) as a part of the VP.

"The vacation packages include hotel accommodations and, in some cases, tours"; *"Some shows just don't impress, he says, and this is one of them"*.

Consider the coordinated PP phrase in *"Alice traveled [both inside and outside the galaxy]$_{PP}$."* Here, *inside* and *outside* are the conjuncts, *both* is a marker, and *"the galaxy"* is a shared argument. A good representation of the coordination structure would allow us to identify the different elements and their associated functions. As we show below, it is often not possible to reliably extract such information from the existing PTB annotation scheme.

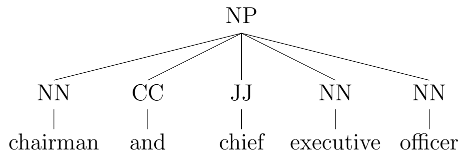
3 Coordinations in the Penn Tree Bank

We now turn to describe how coordination is handled in the PTB, focusing on the parts where we find the annotation scheme to be deficient.

There is no explicit annotation for coordination phrases Some coordinators do not introduce a coordination structure. For example, the coordinator *"and"* can be a discourse marker connecting two sentences (e.g. *"And they will even serve it themselves"*), or introduce a parenthetical (e.g. *"The Wall Street Journal is an excellent publication that I enjoy reading (and must read) daily"*). These are not explicitly differentiate in the PTB from the case where *"and"* connects between at least two elements (e.g. *"loyalty and trust"*).

NPs without internal structure The PTB guidelines (Bies et al., 1995) avoid giving any structure to NPs with nominal modifiers. Following this, 4759 NPs that include coordination were left flat, i.e. all the words in the phrase are at the same level. For example *(NP (NNP chairman) (CC and) (NP chief executive officer))* which is annotated in the PTB as:

[1]



It is impossible to reliably extract conjunct boundaries from such structures. Although work has been done for giving internal structures for flat NPs (Vadas and Curran, 2007), only 48% of the flat NP coordinators that include more than two nouns were given an internal structure, leaving 1744 cases of flat NPs with ambiguous conjunct boundaries.

Coordination parts are not categorized Coordination phrases may include markers, shared modifiers, shared arguments, connective adjectives and parentheticals. Such elements are annotated on the same level as the conjuncts[4]. This is true not only in the case of flat NPs but also in cases where the coordination phrase elements do have internal structures. For examples:

- The *Both* marker in *(NP (DT both) (NP the self) (CC and) (NP the audience))*
- The parenthetical *maybe* in *(NP (NP predictive tests) (CC and) (PRN , maybe ,) (NP new therapies))*
- The shared-modifier "the economy's" in *(NP (NP the economy's) (NNS ups) (CC and) (NNS downs))*

Automatic categorization of the phrases elements is not trivial. Consider the coordination phrase *"a phone, a job, and even into a school"*, which is annotated in the PTB where the NPs *"a phone"* and *"a job"*, the ADVP *"even"* and the PP *"into a school"* are siblings. A human reader can easily deduce that the conjuncts are *"a phone"*, *"a job"* and *"into a school"*, while *"even"* is a connective. However, for an automatic analyzer, this structure is ambiguous: NPs can be conjoined with ADVPs as well as PPs, and a coordination phrase of the form NP NP CC ADVP PP has at least two possible interpretations: (1) Coord Coord CC Conn Coord (2) Coord Coord CC Coord Shared.

Inconsistency in shared elements and markers level The PTB guidelines allows inconsistency in the case of shared ADVP pre-modifiers of VPs (e.g. *"deliberately chewed and winked"*). The pre-modifier may be annotated in the same level of the VP *((ADVP deliberately) (VP chewed and winked))* or inside it *(VP (ADVP deliberately) chewed and winked))*. In addition to this documented inconsistency, we also found markers that are inconsistently annotated in and outside the coordination phrase, such as *respectively* which is

[4] shared arguments may appear in the PTB outside the coordination phrase. For example *He* is an argument for *bought* and for *sold* in *((He) ((bought) (and) (sold) (stocks)))*.

tagged as sibling to the conjuncts in *(NP (NP Feb. 1 1990) (CC and) (NP May. 3 1990), (ADVP respectively))* and as sibling to the conjuncts parent in *(VP (VBD were) (NP 7.37% and 7.42%), (ADVP respectively)).*

Inconsistency in comparative quantity coordination Quantity phrases with a second conjunct of *more*, *less*, *so*, *two* and *up* are inconsistently tagged. Consider the following sentences: *"[50] [or] [so] projects are locked up"*, *"Street estimates of [$ 1] [or so] are low"*. The coordination phrase is similar in both the sentences but is annotated differently.

Various errors The PTB coordination structures include errors. Some are related to flat coordinations (Hogan, 2007). In addition, we found cases where a conjunct is not annotated as a complete phrase, but with two sequenced phrases. For instance, the conjuncts in the sentence *"But less than two years later, the LDP started to crumble, and dissent rose to unprecedented heights"* are *"the LDP started to crumble"* and *"dissent rose to unprecedented heights"*. In the PTB, this sentence is annotated where the first conjunct is splitted into two phrases: *"[the LDP] [started to crumble], and [dissent rose to unprecedented heights]"*.

4 Extended Coordination Annotation

The PTB annotation of coordinations makes it difficult to identify phrases containing coordination and to distinguish the conjuncts from the other parts of a coordination phrase. In addition it contains various errors, inconsistencies and coordination phrases with no internal structure. We propose an improved representation which aims to solve these problems, while keeping the deviation from the original PTB trees to a minimum.

4.1 Explicit Function Marking

We add function labels to non-terminal symbols of nodes participating in coordination structures. The function labels are indicated by appending a -XXX suffix to the non-terminal symbol, where the XXX mark the function of the node. Phrases containing a coordination are marked with a CCP label. Nodes directly dominated by a CCP node are assigned one of the following labels according to their function: *CC* for coordinators, *COORD* for conjuncts, *MARK* for markers[5], *CONN* for connectives and parentheticals, and *SHARED* for shared modifiers/arguments. For shared elements, we deal only with those that are inside the coordination phrase. We do not assign function labels to punctuation symbols and empty elements. For example, our annotation for the sentence *"... he observed among his fellow students and, more important, among his officers and instructors ... "* is:

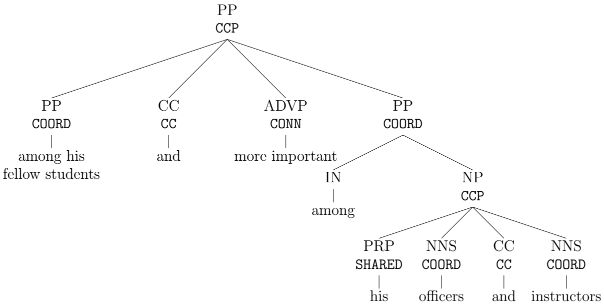


Table 1 summarizes the number of labels for each type in the enhanced version of the Penn Treebank.

Function label	#
CC	24,572
CCP	24,450
COORD	52,512
SHARED	3372
CONN	526
MARK	522

Table 1: The number of labels that were added to the Penn Treebank by type.

4.2 Changes in Tree Structure

As a guiding principle, we try not to change the structure of the original PTB trees. The exceptions to this rule are cases where the structure is changed to provide internal structure when it is missing, as well as when fixing systematic inconsistencies and occasional errors.

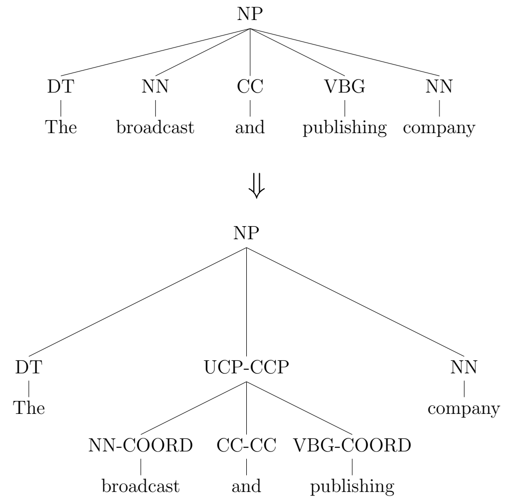
1. In flat coordination structures which include elements with more than one word, we add brackets to delimit the element spans. For instance, in the flat NP in [1] we add brackets to delimit the conjunct *"chief executive officer"*. The full phrase

[5] *both*, *either*, *between*, *first*, *neither*, *not*, *not only*, *respectively* and *together*

structure is: *(NP-CCP (NN-COORD chairman) (CC-CC and) (NP-COORD chief executive officer))*.

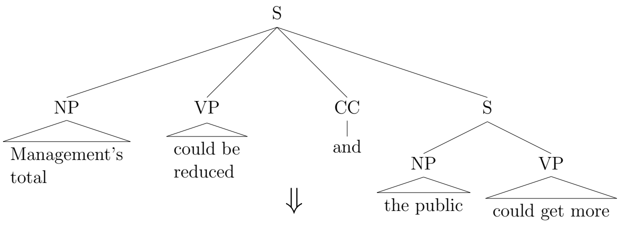
2. Comparative quantity phrases (*“5 dollars or less”*) are inconsistently analyzed in the PTB. When needed, we add an extra bracket with a QP label so they are consistently analyzed as *“5 dollars [or less]$_{QP}$”*. Note that we do not consider these cases as coordination phrases.

3. We add brackets to delimit the coordination phrase in flat cases that include coordination between modifiers while the head is annotated in the same phrase:



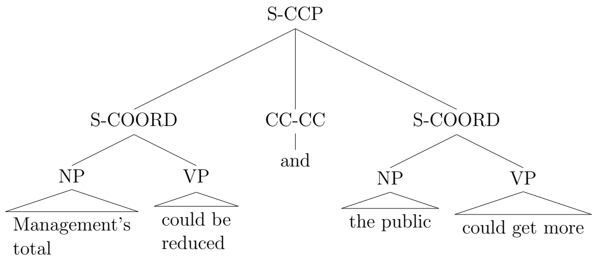
company, which is the head of the phrase, is originally annotated at the same level as the conjuncts *broadcast* and *publishing*, and the determiner *the*. In such cases, the determiner and modifiers are related to the head which is not part of the coordination phrase, requiring the extra bracketing level to delimit the coordination. This is in contrast to the case of coordination between verbs (e.g *“Bob (VP cleaned and refueled the spaceship)”*), where the non coordinated elements (*“the spaceship”*) are shared.

4. When a conjunct is split into two phrases or more due to an error, we add extra brackets to delimit the conjunct as a complete phrase:



Type	#
(1) Flat structures	1872
(2) Comparative quantity phrases	52
(3) Coordination between modifiers	1264
(4) Coordination with errors	213
(5) ADVP inconsistency	206

Table 2: The number of subtrees in the Penn Treebank that were changed in our annotation by type.



5. We consolidate cases where markers and ADVP pre-modifiers are annotated outside the coordination phrase, so they are consistently annotated inside the coordination phrase.

Table 2 summarizes the numbers and types of subtrees that receive a new tree structure in the enhanced version of the Penn Treebank.

5 The Annotation Process

Some of the changes can be done automatically, while other require human judgment. Our annotation procedure combines automatic rules and manual annotation that was performed by a dedicated annotator that was trained for this purpose.

5.1 Explicit marking of coordination phrases

We automatically annotate coordination phrases with a CCP function label. We consider a phrase as coordination phrase if it includes a coordinator and at least one phrase on each side of the coordinator, unlike coordinators that function as discourse markers or introduce parentheticals, which appear as the first element in the phrase.

5.2 Assigning internal structure to flat coordinations

Flat coordinations that include only a coordinator and two conjuncts (e.g. *(NP (NNP Poland) (CC and) (NNP Hungary))*) are trivial and are left with the same structure. For the rest of the flat coordinations (3498 cases), we manually annotated the elements spans. For example, given the flat

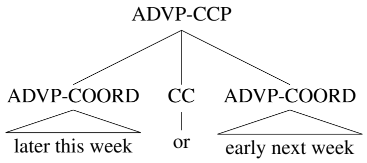
NP: *"[General]$_{NNP}$ [Electric]$_{NNP}$ [Co.]$_{NNP}$ [executives]$_{NNS}$ [and]$_{CC}$ [lawyers]$_{NNS}$"*. The annotator is expected to provide the analysis: *"[General Electric Co.] [executives] [and] [lawyers]"*. We then add brackets around multi-token elements (e.g. *"General Electric Co."*), and set the label according the syntactic structure. The annotation was done while ignoring inner structures that were given in the NP-Bracketing extension of Vadas and Curran (2007). We compare agreement with their annotations in the next section.

To handle cases such as in 4.2(3), where the coordination is between modifiers of a head which is annotated in the PTB on the same level of the conjuncts, we first identify potential candidate phrases of this type by looking for coordination phrases where the last element was not tagged by the annotator as a conjunct. Out of this set, we remove cases where we can reliably identify the non-conjunct element as a marker. For the rest of the cases, we distinguish between NP phrases and non-NP phrases. For NP phrases, we automatically add extra brackets to delimit the coordination phrase span so that it includes only the coordinated modifiers. For the rest of the phrases we found that an such automatic procedure was not feasible (consider the ADVP phrases: *(ADVP (RBR farther) (CC and) (RBR farther) (RB apart))* ; *(ADVP (RB up) (CC and) (RB down) (NP (NNP Florida)))*. The first phrase head is *apart* while in the second phrase, *Florida* is a complement). We manually annotated the coordination phrase boundary in these cases.

When adding an extra tree level in this cases, we set its syntactic label to UCP when the conjuncts are from different types and same as the conjuncts label when the conjuncts are from the same type.[6]

5.3 Annotating roles within coordination phrases

Cases where there are only a coordinator and two siblings in the coordinated phrase are trivial to automatically annotate, marking both siblings as conjuncts:



To categorize the phrase elements for the rest of the coordination phrases, we first manually marked the conjuncts in the sentence (for flat structures, the conjuncts were already annotated in the internal structure annotation phase). The annotator was given a sentence where the coordinator and the coordination phrase boundaries are marked. For example "`Coke has been able to improve (bottlers' efficiency and production, {and} in some cases, marketing)`". The annotation task was to mark the conjuncts.[7] We automatically concluded the types of the other elements according to their relative position – elements before or after the conjuncts are categorized as markers/shared, while an element between conjuncts is a connective or the coordinator itself.

Mismatches with the PTB phrase boundaries In 5% of the cases of coordination with inner structure, a conjunct span as it was annotated by our annotator was not consistent with the elements spans in the PTB. For example, the annotator provided the following annotation: "`(The [economic loss], [jobs lost], [anguish],[frustration] {and} [humiliation]) are beyond measure`", treating the determiner "The" as a shared modifier. In contrast, the PTB analysis considers "The" as part of the first conjunct ("`[The economic loss]`").

The vast majority of the mismatches were on the point of a specific word such as *the* (as demonstrated in the above example), *to*, *a* and punctuation symbols. In a small number of cases the mismatch was because of an ambiguity. For example, in "`The declaration immediately made the counties eligible for (temporary housing, grants {and} low-cost loans to cover uninsured property losses)`" the annotator marked *"temporary housing"*, *"grants"*, and *"low-cost loans"* as conjuncts (leaving *"to cover uninsured property loss"* as a shared

[6] When the conjuncts are in POS level, a corresponding syntactic label is set. For example: ***(NP-CCP** (NN-COORD head) (CC-CC and) (NNS-COORD shoulders))*

[7] The coordination phrase boundaries were taken from the PTB annotations and were used to focus the annotators attention, rather than to restrict the annotation. The annotators were allowed to override them if they thought they were erronous. We did not encounter such cases.

modifier, while the PTB annotation considers *"to cover... "* as part of the last conjunct. Following our desiderata of minimizing changes to existing tree structures, in a case of a mismatch we extend the conjunct spans to be consistent with the PTB phrasing (each such case was manually verified).

5.4 Handling inconsistencies and errors

We automatically recognize ADVPs that appear right before a VP coordination phrase and markers that are adjunct to a coordination phrase. We change the structure such that such ADVPs and markers appear inside the coordination phrase.

Quantity phrases that includes two conjuncts with a second conjunct of *more, less, so, two* and *up* are automatically recognized and consolidated by adding an extra level.

Errors in conjuncts span are found during the manual annotation that is done for the categorization. When the manual annotation includes a conjunct that is originally a combination of two siblings phrases, we add extra brackets and name the new level according to the syntactic structure.

6 Annotator Agreement

We evaluate the resulting corpus with inter-annotators agreement for coordination phrases with inner structure as well as agreement with the flat conjuncts that were annotated in the NP bracketing annotation effort of Vadas and Curran (2007).

6.1 Inter-annotator agreement

To test the inter-annotator agreement, we were assisted with an additional linguist who annotated 1000 out of 7823 coordination phrases with inner structure. We measured the number of coordination phrases where the spans are inconsistent at least in one conjunct. The annotators originally agreed in 92.8% of the sentences. After revision, the agreement increased to 98.1%. The disagreements occurred in semantically ambiguous cases. For instance, *"potato salad, baked beans and pudding, plus coffee or iced tea"* was tagged differently by the 2 annotators. One considered *"pudding"* as the last conjunct and the other marked *"pudding, plus coffee or iced tea"*.

6.2 Agreement with NP Bracketing for flat coordinations

The NP Bracketing extension of Vadas and Curran (2007) includes inner structures for flat NP phrases in the PTB, that are given an internal structure using the NML tag. For instance, in *(NP (NNP Air) (NNP Force) (NN contract))*, *"Air Force"* is considered as an independent entity and thus is delimited with the NML tag: *(NP (NML (NNP Air) (NNP Force)) (NN contract))*.

As mentioned, 48% (1655 sentences) of the NP flat coordination were disambiguated in this effort.[8] For these, the agreement on the conjuncts spans with the way they were marked by our annotators is 88%. The disagreements were in cases where a modifier is ambiguous. For examples consider *"luxury"* in *"The luxury airline and casino company"*, *"scientific"* in *"scientific institutions or researchers"* and *"Japanese"* in *"some Japanese government officials and businessmen"*. In cases of disagreement we followed our annotators decisions.[9]

	R	P	F1
PTB + NPB	90.41	86.12	88.21
PTB + NPB + CCP	90.83	91.18	91.01

Table 3: The parser results on section 22.

7 Experiments

We evaluate the impact of the new annotation on the PTB parsing accuracy. We use the state-of-the-art Berkeley parser (Petrov et al., 2006), and compare the original PTB annotations (including Vadas and Curran's base-NP bracketing – **PTB+NPB**) to the coordination annotations in this work (**PTB+NPB+CCP**). We use sections 2-21 for training, and report accuracies on the traditional dev set (section 22). The parse trees are scored using EVALB (Sekine and Collins, 1997).

Structural Changes We start by considering how the changes in tree structures affect the parser performance. We compared the parsing performance when trained and tested on PTB+NPB, to the parsing performance when trained and tested on PTB+NPB+CCP. The new function labels were ignored in both training and testing. The results

[8] We consider a flat NP coordination as disambiguated if it includes a coordinator and two other elements, i.e.: *(NML (NML (NN eye) (NN care)) (CC and) (NML (NN skin) (NN care)))* ; *(NML (NN buy) (CC or) (NN sell))*.

[9] A by-product of this process is a list of ambiguous modifier attachment cases, which can be used for future research on coordination disambiguation, for example in designing error metrics that take such annotator disagreements into account.

Gold \ Pred	**CC**	**CCP**	**COORD**	**MARK**	**SHARED**	**CONN**	**None**	**Err**
CC	849					1	5	
CCP		552	1				91	205
COORD		3	1405		2		184	200
MARK				9			2	1
SHARED	1				29		85	3
CONN						1	4	2
None	4	124	113	4	26	14		

Table 4: Confusion-matrix over the predicted function labels. **None** indicate no function label (a constituent which is not directly inside a CCP phrase). **Err** indicate cases in which the gold span was not predicted by the parser.

are presented in Table 3. Parsing accuracy on the coordination-enhanced corpus is higher than on the original trees. However, the numbers are not strictly comparable, as the test sets contain trees with somewhat different number of constituents. To get a fairer comparison, we also evaluate the parsers on the subset of trees in section 22 whose structures did not change. We check two conditions: trees that include coordination, and trees that do not include coordination. Here, we see a small drop in parsing accuracy when using the new annotation. When trained and tested on PTB+NPB+CCP, the parser results are slightly decreased compared to PTB+NPB – from 89.89% F1 to 89.4% F1 for trees with coordination and from 91.78% F1 to 91.75% F1 for trees without coordination. However, the drop is small and it is clear that the changes did not make the corpus substantially harder to parse. We also note that the parsing results for trees including coordinations are lower than those for trees without coordination, highlighting the challenge in parsing coordination structures.

Function Labels How good is the parser in predicting the function labels, distinguishing between conjuncts, markers, connectives and shared modifiers? When we train and test the parser on trees that include the function labels, we see a rather large drop in accuracy: from 89.89% F1 (for trees that include a coordination) to 85.27% F1. A closer look reveals that a large part of this drop is superficial: taking function labels into account cause errors in coordination scope to be punished multiple times.[10] When we train the parser with function labels but ignore them at evaluation time, the results climb back up to 87.45% F1. Furthermore, looking at coordination phrases whose structure was perfectly predicted (65.09% of the cases), the parser assigned the correct function label for all the coordination parts in 98.91% of the cases. The combined results suggest that while the parser is reasonably effective at assigning the correct function labels, there is still work to be done on this form of disambiguation.

The availability of function labels annotation allows us to take a finer-grained look at the parsing behavior on coordination. Table 4 lists the parser assigned labels against the gold labels. Common cases of error are (1) conjuncts identification – where 200 out of 1794 gold conjuncts were assigned an incorrect span and 113 non-conjunct spans were predicted as participating as conjuncts in a coordination phrase; and (2) Shared elements identification, where 74.57% of the gold shared elements were analyzed as either out of the coordination phrase or as part of the last coordinates. These numbers suggest possible areas of future research with respect to coordination disambiguation which are likely to provide high gains.

8 Conclusions

Coordination is a frequent and important syntactic phenomena, that pose a great challenge to automatic syntactic annotation. Unfortunately, the current state of coordination annotation in the PTB is lacking. We present a version of the PTB with improved annotation for coordination structure. The

[10] Consider the gold structure (NP (NP-CCP (DT-MARK a) (NP-COORD b) (CC and) (NP-COORD c) (PP-SHARED d))) and the incorrect prediction (NP (DT a) (NP-CCP (NP-COORD b) (CC and) (NP-COORD c)) (PP d)). When taking only the syntactic labels into account there is only the mistake of the coordination span. When taking the coordination roles into account, there are two additional mistakes – the missing labels for a and d.

new annotation adds structure to the previously flat NPs, unifies inconsistencies, fix errors, and marks the role of different participants in the coordination structure with respect to the coordination. We make our annotation available to the NLP community. This resource is a necessary first step towards better disambiguation of coordination structures in syntactic parsers.

Acknowledgments

This work was supported by The Allen Institute for Artificial Intelligence as well as the German Research Foundation via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. *University of Pennsylvania*, 97:100.

David Dowty. 1988. Type raising, functional composition, and non-constituent conjunction. In *Categorial grammars and natural language structures*, pages 153–197. Springer.

Kazuo Hara, Masashi Shimbo, Hideharu Okuma, and Yuji Matsumoto. 2009. Coordinate structure analysis with global structural constraints and alignment-based local features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 967–975. Association for Computational Linguistics.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. Association for Computational Linguistics.

Rodney Huddleston, Geoffrey K Pullum, et al. 2002. The cambridge grammar of english. *Language. Cambridge: Cambridge University Press*, pages 1273–1362.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Wolfgang Maier, Erhard Hinrichs, Sandra Kübler, and Julia Krivanek. 2012. Annotating coordination in the penn treebank. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 166–174. Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

Satoshi Sekine and Michael Collins. 1997. Evalb bracket scoring program. *URL http://nlp. cs. nyu. edu/evalb/EVALB. tgz.*

Masashi Shimbo and Kazuo Hara. 2007. A discriminative learning model for coordinate conjunctions. In *EMNLP-CoNLL*, pages 610–619.

David Vadas and James Curran. 2007. Adding noun phrase structure to the penn treebank. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 240.