A Corpus-based Syntactic Analysis of Two-termed Unlike Coordination

Julie Kallini and **Christiane Fellbaum**
Department of Computer Science
Princeton University
{jkallini,fellbaum}@princeton.edu

Abstract

Coordination is a phenomenon of language that conjoins two or more terms or phrases using a coordinating conjunction. Although coordination has been explored extensively in the linguistics literature, the rules and constraints that govern its structure are still largely elusive and widely debated amongst linguists. This paper presents a study of two-termed *unlike* coordinations in particular, where the two conjuncts of the coordination phrase form valid constituents but have distinct categories. We conducted a syntactic analysis of the phrasal categories that can be conjoined in such unlike coordinations through a computational corpus-based approach, utilizing the Corpus of Contemporary American English (COCA) as the main data source, as well as the Penn Treebank (PTB). The results show that the two conjuncts within unlike coordinations display different properties based on their position, supporting an antisymmetric view of the structure of coordination. This research provides new data and perspectives through the use of statistical techniques that can help shape future theories and models of coordination.

1 Introduction

1.1 Motivation

Coordination is a phenomenon of language that conjoins two or more terms or phrases. The terms or phrases that are grouped in coordination phrases are normally called *conjuncts*, and they are often conjoined by a *coordinating conjunction*, such as *and*, *or*, *but*, or *nor*. A common assumption in the linguistics literature is that two elements may only be coordinated if they share the same syntactic category, as in (1).

(1) a. [$_{NP}$ The chicken] and [$_{NP}$ the rice] go well together.

b. The president will [$_{VP}$ understand the criticism] and [$_{VP}$ take action].

For example, in (1a), the two conjuncts being coordinated are "the chicken" and "the rice," which share the same syntactic category of *noun phrase* (NP). The assumption that the conjuncts of a coordination phrase will always have the same category is known as the Law of the Coordination of Likes (LCL) (Williams, 1981). The LCL explains why many instances of coordination are ungrammatical, such as the coordination of a prepositional phrase (PP) and a clause (CP) shown in (2) (Prażmowska, 2015).

(2) a. The scene of the movie was in Chicago.

b. The scene that I wrote was in Chicago.

c. * The scene [$_{PP}$ of the movie] and [$_{CP}$ that I wrote] was in Chicago.

Even though the prepositional phrase and the clause are both grammatical when standing alone within the context sentence, as in (2a) and (2b), their coordination in (2c) is ungrammatical, supposedly because of the LCL.

However, several examples of syntactically *unlike* coordination can be found in English, such as the examples in (3) (Sag et al., 1985).

(3) a. Pat is [$_{NP}$ a Republican] and [$_{AP}$ proud of it].

b. John is [$_{AP}$ healthy] and [$_{PP}$ in good shape].

c. That was [$_{NP}$ a rude remark] and [$_{PP}$ in very bad taste].

In the above examples, the two conjuncts within each coordination phrase do not share the same syntactic category. In these cases, the LCL seems to be too restrictive.

Yet, there are also cases in which the LCL is not restrictive enough—a coordination phrase can still be ungrammatical even if its conjuncts have the same syntactic category (Prażmowska, 2015).

(4) a. * John ate [$_{PP}$ with his mother] and [$_{PP}$ with good appetite].

b. * John [$_{AdvP}$ probably] and [$_{AdvP}$ unwillingly] went to bed.

Example (4a) contains the coordination of two prepositional phrases, and (4b) contains the coordination of two adverbs. Despite the two conjuncts having like categories, these examples result in ungrammatical sentences. Semantics seems to play a role in the acceptability of coordinations as well; a stronger version of the LCL requires that conjuncts must also be alike in their semantic function. For example, in (4a), the first prepositional phrase "with his mother" expresses accompaniment, whereas the second "with good appetite" expresses manner (Prażmowska, 2015). However, identifying and articulating rigorous rules that predict all grammatical possibilities of coordination has been a difficult task for linguists, and as a result, the underlying syntactic structure of coordination phrases has been elusive.

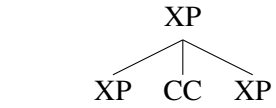
1.2 Goal

The goal of this project is to explore and answer questions about the syntax of coordination phrases through a quantitative corpus analysis. By analyzing a large corpus of naturally-occurring spoken and written language using natural language processing and statistical techniques, we will investigate the patterns of syntactic categories found in unlike coordinations. An overarching goal for this project is to share data that may inform linguistic hypotheses about the underlying structure of coordination.

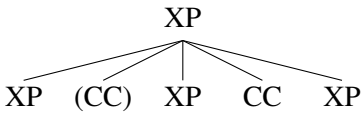
By taking a computational approach, we can explore a larger and deeper set of questions regarding coordination, such as:

- What combinations of syntactic categories are attested in English data, and which appear most frequently?
- Does this depend on the genre of the text or the type of conjunction (*and*, *or*, *but*, *nor*)?

This paper begins by introducing the relevant problem background and related work. We then detail our corpus-based approach and implementation, which utilizes the Corpus of Contemporary American English (COCA), the Penn Treebank (PTB), and the Berkeley Neural Parser. We then follow with a presentation of the results and provide an in-depth discussion of the significant findings.



(a) Two-termed coordination.



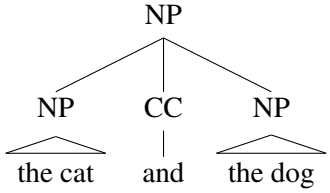
(b) Multiple-termed coordination.

Figure 1: Flat multi-headed proposal for the structure of coordination.

2 Background and Related Work

Capturing the structure of coordination has been a difficult problem in many theories of syntax. A flat, multi-headed structure was proposed in earlier theories, in which two or more lexical heads share the same phrase-level projection, as in the templates shown in Figure 1 (Progovac, 1998a; Chomsky, 1981). This theory captures the intuitive idea that the coordination of two NPs is an NP, that the coordination of two VPs is a VP, etc. An example of a two-termed coordination of NPs is provided in (5). We use *CC* as the name for the functional category of coordinating conjunctions, which is also the label used in the PTB.

(5) [$_{NP}$ the cat] and [$_{NP}$ the dog]



There are several problems with this view, but the problem we are most concerned with relates to the aforementioned counterexamples to the LCL, restated below in (6).

(6) a. Pat is [$_{NP}$ a Republican] and [$_{AP}$ proud of it].

b. John is [$_{AP}$ healthy] and [$_{PP}$ in good shape].

c. That was [$_{NP}$ a rude remark] and [$_{PP}$ in very bad taste].

In fact, the LCL was formulated due to this proposal for the syntax of coordination. Coordination was said to denote a relation between two (or more) elements that are "hierarchically equal" in that neither of the elements is more prominent than the

other, leading to a symmetrical and flat vision of coordination structures (Prażmowska, 2015). Since conjuncts were assumed to be symmetrical and equal in status, it followed that they must share the same syntactic category to be grammatically coordinated.

One proposal that seems to address the existence of the unlike category coordinations seen in (6) is Bowers's *Pred* (predicate) functional category (Bowers, 1993). On top of the NPs, APs, and PPs being coordinated in these sentences, there is another level of structure. Bowers suggests that a null Pred head selects an NP, AP, or PP as its complement, forming a predicate phrase (PredP). Thus, unlike coordinations are actually like coordinations in disguise—all conjuncts have the category of PredP. PredPs are complements of the copula *be* in these sentences, as made apparent in (7).

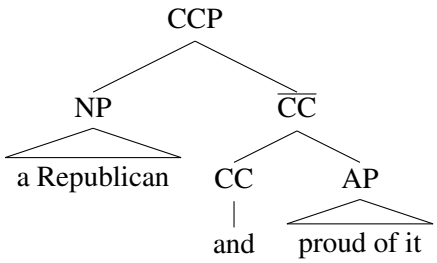
(7) a. Pat is [$_{\text{PredP}}$ Ø [$_{\text{NP}}$ a Republican]] and [$_{\text{PredP}}$ Ø [$_{\text{AP}}$ proud of it]].
b. John is [$_{\text{PredP}}$ Ø [$_{\text{AP}}$ healthy]] and [$_{\text{PredP}}$ Ø [$_{\text{PP}}$ in good shape]].
c. That was [$_{\text{PredP}}$ Ø [$_{\text{NP}}$ a rude remark]] and [$_{\text{PredP}}$ Ø [$_{\text{PP}}$ in very bad taste]].

However, Bowers's proposal does not account for cases where the coordinated strings are not predicates, such as in (8). In each of these examples, the coordination phrase is an adjunct of VP rather than a predicate complement of VP, and the conjuncts semantically serve the purpose of adverbial modification.

(8) a. The surgeon operated [$_{\text{AdvP}}$ slowly] and [$_{\text{PP}}$ with great care].
b. Alice will visit home [$_{\text{AdvP}}$ tomorrow] or [$_{\text{PP}}$ on the weekend].

Other proposals dodge the problem of unlike coordination entirely by making the coordinating conjunction the head of its own coordination phrase (*CCP*). One example of such a theory is shown in (9).

(9) [$_{\text{NP}}$ a Republican] and [$_{\text{AP}}$ proud of it]



Here, conjuncts are specifiers and complements of the head conjunction (Johannessen, 1998; Zoerner, 1995). With such a construction, the categories of the conjuncts by themselves do not pose a restriction on the possibility of coordination. Thus, such theories do not have anything to say about the LCL, but they are still problematic in that they over-generate; no combinations of categories are prohibited.

3 Approach

We approached the task of capturing the structure of two-termed coordination by conducting a computational syntactic analysis on a large quantity of corpus data. Our primary data source is the Corpus of Contemporary American English (COCA) (Davies, 2015), and our additional data source is the Penn Treebank (PTB) augmented with Ficler and Goldberg's PTB coordination annotation extension (Ficler and Goldberg, 2016). We extracted coordination phrases from both of these datasets and performed a quantitative syntactic analysis using the constituency parses of the sentences within both texts.

This approach has a few advantages over previous work. Much of the research that has shaped current theories of coordination have relied on the acceptability judgments of a few individuals, usually the author(s). By using corpus data, we gain an understanding of coordination on a much larger scale and emphasize empirical rather than intuitive judgments. We can also investigate differences in the patterns we identify based on the genre from which a coordination was found or the conjunction it contains.

3.1 Corpus Data

The Corpus of Contemporary American English (COCA) is a large, genre-balanced corpus of American English containing more than 450 million words of text (Davies, 2015). The COCA contains text from five genres: academic, fiction, magazine, newspaper, and spoken texts. Each genre includes 20 million words each year from 1990-2012. A balanced corpus, especially one that includes spoken data, was important for this project, as there may be variations in the coordinations found across different genres.

In addition to COCA data, we use the Penn Treebank (PTB), a collection of 2,499 stories from the Wall Street Journal gathered over a three-year pe-

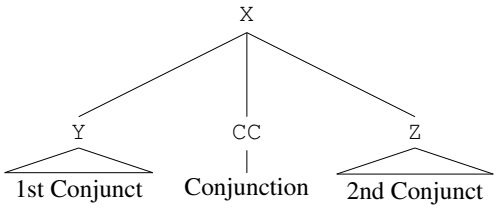
riod (Marcus et al., 1993). Sentences from the PTB are already tokenized and annotated with phrase structure, unlike the COCA. However, coordination annotations in the PTB are often inconsistent, include errors, and lack internal structure in many cases. For this reason, we make use of Ficler and Goldberg's PTB coordination annotation extension, which improves the coordination annotation in the PTB (Ficler and Goldberg, 2016). This extension provides an annotation that explicitly marks coordination phrases and the role of each element in coordination structures (i.e., conjuncts, markers, connectives, and shared elements are all identified and marked).

3.2 Syntactic Analysis

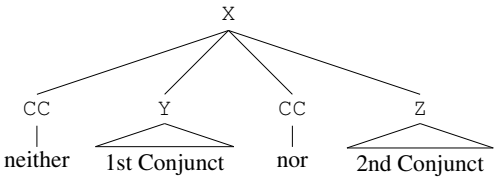
The main task of our syntactic analysis involves the detection and extraction of coordination phrases from our corpus data. Since the COCA is provided in a raw text format, we use the Berkeley Neural Parser to produce syntax trees of sentences in the COCA. This is a state-of-the-art constituency parser that generates syntax trees in the style of the Penn Treebank (Kitaev and Klein, 2018). To implement a good search algorithm for coordinations within parsed COCA data, we studied several sentence parse trees containing coordinations and identified three patterns in the way that the Berkeley Neural Parser most often represents the structure of coordination phrases, as shown in Figure 2.

Since the PTB is already annotated as phrase structure trees, the possible problems of using a constituency parser on novel text are eliminated. The identification of coordination phrases is made much simpler here with the help of the coordination annotation extension. The explicit function markers allow for the straightforward detection and isolation of conjuncts and conjunctions from other tangential elements that may be contained within a coordination phrase, such as modifiers and connectives. Figure 3 shows an example of a PTB phrase structure tree with the extension's additional function marking.

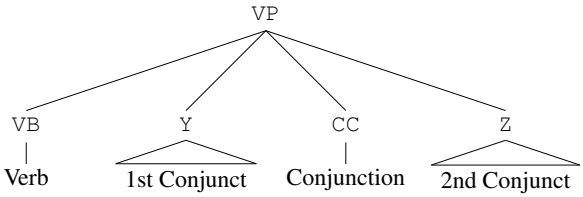
For our syntactic analysis, we include coordinations of six types of PTB phrasal category labels: noun phrases (NP), verb phrases (VP), prepositional phrases (PP), adjective phrases (ADJP), adverb phrases (ADVP), and subordinate clauses (SBAR, often called *complementizer phrases* (CP) in more recent syntax literature). We have chosen this set of labels because they correspond to the most frequent phrasal categories in the data. Once coordination phrases have been identified, we run statistical tests on the frequencies of their different attributes, such as the categories of the conjuncts, the type of conjunction used, and the genre from which the coordination was found.



(a) Simple ternary-branching pattern.



(b) *Neither-nor* pattern.



(c) Verb-complement pattern.

Figure 2: Three patterns used to detect two-termed coordination phrases in parsed COCA data. X, Y, and Z may be any PTB constituent tags.

4 Results

In our analyses, we employ the *chi-square* (χ^2) tests, which determine whether a set of observed frequencies deviate significantly from a set of expected frequencies. We consider p-values less than 0.05 to be statistically significant. Since our sam-

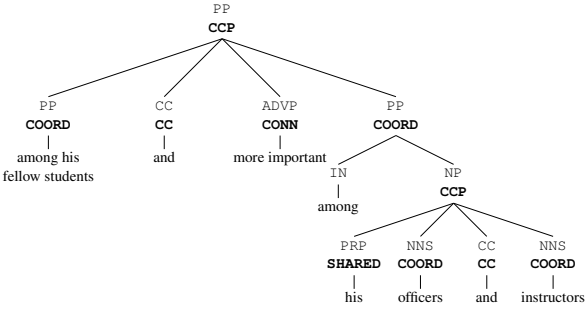


Figure 3: A tree containing the explicit function marking from the PTB coordination annotation extension.

V	Association
0.00–0.05	negligible
0.05–0.10	weak
0.10–0.15	moderate
0.15–0.25	strong
0.25–1.00	very strong

Table 1: Interpretation of strength of association/tendency based on Cramer's V.

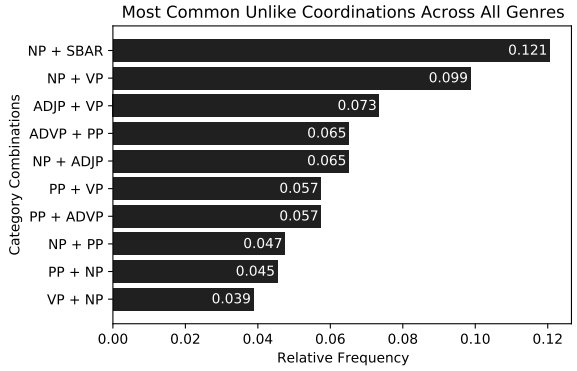


Figure 4: Most frequent unlike category combinations in the COCA data. Frequencies are relative to all unlike coordinations.

ple sizes are very large, we conduct additional post-tests to accompany any statistically significant results. We use Cramer's V to measure strength of association (Table 1) (Akoglu, 2018).

4.1 Most Frequent Unlike Coordinations

We performed an analysis of the most frequent unlike coordinations in the COCA data. Figure 4 displays the top ten most common unlike coordinations found in all of the COCA data we parsed along with their relative frequencies, and Table 2 contains examples. We found a significant difference in the distribution of unlike category coordinations, with a moderate tendency toward the most common coordination combination, NP+SBAR, $\chi^2(9, N = 24456) = 3142.0$, $p < .001$, $V = .119$.

4.1.1 By COCA Genre

We also performed an analysis of the most frequent unlike coordinations in each of the five COCA genres. In each genre, a significant difference was found in the distribution of unlike category coordinations. Table 3 summarizes the results of the chi-square tests and Cramer's V for each COCA genre, and Appendix B contains figures displaying the top

Coordination	Example Sentence
NP+SBAR*	You'd get to watch two adults talk about [NP America] and [SBAR what they would do to lead it].
NP+VP	Voids are [NP a nightmare] and [VP initialed by the employee and his supervisor].
ADJP+VP*	It was [ADJP emotionally manipulative] and [VP designed to scare people into faith].
ADVP+PP*	The phenomenon fell into place [ADVP organically] and [PP with ease].
NP+ADJP*	He's [NP a free spirit] and [ADJP playful], prompting managers and teammates to shake their heads and proclaim he's Manny being Manny.
PP+VP	In Gaza, meanwhile, Hamas leaders insist that they are still [PP in charge] and [VP leading the Palestinian authority].
PP+ADVP*	A big question many taxpayers face is whether to file [PP by paper] or [ADVP electronically].
NP+PP	I called him a liar again, and then I punched him [NP a lot of times] and [PP with all my might].
PP+NP	More Americans work [PP out of the house] and [NP longer hours], so we've become more dependent on meals we don't cook ourselves.
VP+NP	Erosion and years of neglect have left the brick structure [VP crumbling] and [NP a clear safety hazard].

* Also in the top ten unlike coordinations in the PTB.

Table 2: Examples extracted from the COCA for each of the top ten most common unlike coordinations.

Genre	χ^2	N	p	V
Academic	450.59	5105	$< .001$	.099
Fiction	693.22	4358	$< .001$	.133
Magazine	583.69	5324	$< .001$	.110
Newspaper	616.91	4851	$< .001$	.118
Spoken	2391.3	5095	$< .001$	.228

Table 3: Summary of chi-square test and Cramer's V results for the frequency difference among the top ten unlike category combinations in each COCA genre.

unlike coordinations in each genre. In the academic genre, there was a weak tendency toward the most common combination, NP+SBAR (Figure 7). In the fiction genre, a moderate tendency was found toward the most common combination, ADJP+VP (Figure 8). In the magazine genre, we also found a moderate tendency toward the most common combination, which was again NP+SBAR, as in the academic genre (Figure 9). In the newspaper genre, an indication of a moderate tendency toward the most common combination was found once again, with NP+VP being the most common combination (Figure 10). In the spoken genre, there is a notable indication of a strong tendency toward the most common combination, which was NP+SBAR, as in the academic and magazine genres (Figure 11).

Conjunction	χ^2	N	p	V
and	2933.0	19621	$< .001$	.129
or	752.87	4317	$< .001$	.139
but	73.893	1042	$< .001$	.089
nor	14.333	45	.111	-

Table 4: Summary of chi-square test and Cramer's V results for the frequency difference among the most common unlike coordinations based on the coordinating conjunction used to conjoin them (from COCA data).

4.1.2 By Conjunction

We also performed an analysis of the most frequent unlike category combinations based on the type of coordinating conjunction used to conjoin them. Table 4 summarizes the results of the chi-square tests and Cramer's V for each type of conjunction, and Appendix B again contains figures displaying the top unlike coordinations for each conjunction. For the conjunctions *and*, *or*, and *but*, a significant difference was found in the distribution of unlike category coordinations. For unlike coordinations containing *and*, there was a moderate tendency toward the most common combination, which was NP+SBAR (Figure 12). For unlike coordinations containing *or*, we also found a moderate tendency toward the most common combination, which was again NP+SBAR (Figure 13). For unlike coordinations containing *but*, there was a weak tendency toward the most common combination, ADJP+VP (Figure 14). For unlike coordinations containing *nor*, no significant difference was found in the distribution of unlike category coordinations (Figure 15).

4.1.3 In the PTB

We performed an analysis of the most frequent unlike coordinations in the PTB as well. Figure 5 displays the top ten most common unlike coordinations in the PTB data, along with their relative frequencies. We found a significant difference in the distribution of unlike category coordinations with a moderate tendency toward the most common combinations, $\chi^2(9, N = 216) = 22.981$, $p = .006$, $V = .109$. The most common unlike coordination in the PTB was ADVP+PP.

4.2 Differences Between Conjunct Positions

In addition to the most frequent combinations of categories, we conducted an analysis of the categories for each conjunct independently. We first report the results based on frequencies from the COCA. Table 5 summarizes the results of the chi-square tests and Cramer's V for each of the six phrasal categories. For NPs, a very strong tendency was found toward the first conjunct position; for VPs, a very strong tendency was found toward the second conjunct position; for PPs, only a weak tendency was found toward the first conjunct position; for ADJPs, a moderate tendency was found toward the first conjunct position; for ADVPs, only a negligible tendency was found toward the second conjunct position; and for SBARs, a very strong tendency was found toward the second conjunct position.

Next, we report the results based on frequencies from the PTB. Table 6 summarizes the results of the chi-square tests and Cramer's V for each of the six phrasal categories. For NPs, a very strong tendency was found toward the first conjunct position; for VPs, PPs, ADJPs, and ADVPs, no significant difference was found in the distribution of conjunct positions; and for SBARs, a very strong tendency was found toward the second conjunct position.

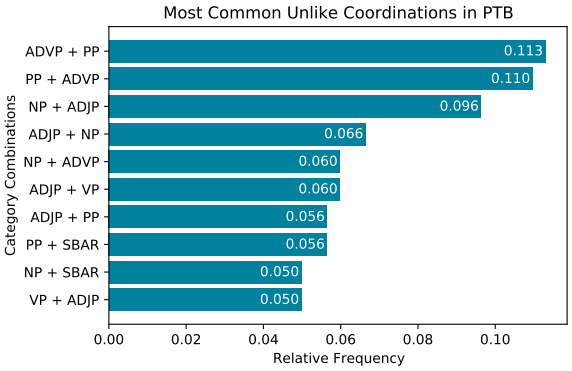


Figure 5: Most frequent unlike category combinations in the PTB. Frequencies are relative to all unlike coordinations.

Category	1st Conjunct	2nd Conjunct	χ^2	N	p	V
NP	70.75%	29.24%	3200.7	18582	$< .001$	.415
VP	32.42%	67.58%	1764.4	14277	$< .001$	.352
PP	53.47%	46.53%	68.789	14248	$< .001$	.069
ADJP	55.73%	44.27%	125.57	9566	$< .001$	.114
ADVP	48.71%	51.29%	5.076	7645	.024	.026
SBAR	23.97%	76.03%	2385.8	8800	$< .001$	.521

Table 5: Summary of chi-square test and Cramer's V results for the frequency difference between the two conjunct positions for each type of phrasal category from COCA data.

Category	1st Conjunct	2nd Conjunct	χ^2	N	p	V
NP	65.57%	34.43%	11.836	122	$< .001$	.311
VP	38.18%	61.82%	3.073	55	.080	-
PP	50.34%	49.66%	.0069	145	.934	-
ADJP	54.17%	45.83%	.8333	120	.361	-
ADVP	44.25%	55.75%	1.496	113	.221	-
SBAR	26.09%	73.91%	10.522	46	.001	.478

Table 6: Summary of chi-square test and Cramer's V results for the frequency difference between the two conjunct positions for each type of phrasal category from the PTB data.

5 Evaluation

A portion of the data we have presented in the previous section was gathered through the use of a constituency parser to identify coordination phrases. While the Berkeley Neural Parser is state-of-the-art, no parser is perfect, especially concerning coordination disambiguation. Furthermore, there are additional types of coordination structures that we do not consider, including non-constituent coordination and gapping. In *non-constituent coordination*, each conjunct in a coordination phrase does not form its own constituent under traditional theories of clause structure, as shown in example (10).

(10) The girl from California walked [into the room at 9 PM] and [out of the room at 10 PM].

Gapping is the phenomenon in which a phrase is coordinated with another phrase that seems to be missing some material, as shown in (11).

(11) [Mary ate beans] and [John ___ potatoes].

While this paper only seeks to analyze the coordination of constituents and does not consider these additional types of coordination, they still pose challenges in the identification and labeling of coordination phrases by parsers. We have conducted an evaluation plan in which human raters manually assessed a random sample of unlike coordinations to estimate an error rate for each type of category combination.

Each type of unlike coordination was assigned a score based on the judgments of three independent raters. A single rater contributes to the score by providing the percentage of samples in which they agreed with the parser's labels. The overall score for that type of coordination is then assigned by taking the mean of the three raters' scores. The scores for each type of unlike coordination are enumerated in Table 7, along with the sample size, confidence level, and margin of error used for sampling.

	NP	VP	PP	ADJP	ADVP	SBAR
NP	-	50.9% (103)	72.3% (100)	72.4% (101)	61.2% (96)	83.4% (103)
VP	61.7% (99)	-	69.0% (96)	70.3% (95)	62.0% (85)	63.0% (105)
PP	61.7% (100)	64.4% (101)	-	80.0% (90)	80.3% (101)	70.7% (97)
ADJP	77.6% (97)	80.0% (102)	89.6% (97)	-	66.6% (78)	65.5% (64)
ADVP	55.5% (86)	66.0% (89)	85.3% (101)	56.5% (76)	-	63.7% (96)
SBAR	79.5% (96)	50.0% (93)	75.3% (81)	58.0% (36)	56.5% (46)	-

Table 7: Average agreement with the Berkeley Neural Parser's labeling of each type of unlike coordination phrase, based on the judgments of the three raters. Rows correspond to the first conjunct's category, and columns correspond to the second conjunct's category. A 90% confidence level and $\pm 8\%$ margin of error were used for sampling each category combination. The sample size, n, is reported in parentheses.

κ	Agreement
0.00–0.20	poor
0.20–0.40	fair
0.40–0.60	moderate
0.60–0.80	substantial
0.80–1.00	near perfect

Table 8: Interpretation of strength of agreement based on the Cohen's Kappa Coefficient.

Fleiss' Kappa showed that, among the three raters, there was fair agreement in their judgments, $\kappa = .291$ (95% CI [.271, .312]), $p < .001$. The strength of agreement is determined based on the Cohen's Kappa Coefficient (Table 8) (McHugh, 2015).

6 Discussion

6.1 Most Frequent Unlike Coordinations

The results of the analysis of the most frequent unlike coordinations in the COCA data indicate that NP+SBAR is the most common unlike coordination. It was also the most frequent unlike coordination in three of the five genres (academic, magazine, and spoken). Some examples from the COCA are shown in (12) below.

(12) a. Be sure to tell us [NP your full name] and [SBAR where you live].

b. I support [NP the president] and [SBAR what he did].

c. The zone's size depends on [NP the weather] and [SBAR how much flow the Mississippi brings each year].

One possible explanation for the high frequency of NP+SBAR coordinations is that subordinate clauses have very similar syntactic distributions to noun phrases in other contexts as well. In particular, subordinate clauses, which are called *complementizer phrases* (CP) in the syntax literature, can be the subjects of sentences. When a CP occupies the subject position of a sentence, it is called a *sentential subject* (Lohndal, 2014). Sentential subjects can be headed by a variety of different complementizers; (13) shows a few examples using *that*, *whether*, *what*, and *how*.

(13) a. [CP That Joe fell asleep in the meeting] disappointed us.
b. [CP Whether she shows up or not] doesn't matter.
c. [CP What a huge scandal it was] didn't emerge until later.
d. [CP How we got here] is a total mystery.

Some linguists have theorized that sentential subjects and more typical nominal subjects have the same syntactic category. Much like Bowers's predicate phrase analysis discussed in Section 2, sentential subjects may be analyzed as having a null determiner head that forms a *determiner phrase* (DP) from a CP (Lohndal, 2014).

(14) [DP Ø [CP That Mary left early]] disappointed us.

Although we do not expound on the arguments for DPs, most of the constituents that we have been treating as noun phrases for simplicity in this paper are often analyzed as DPs instead. The heads of determiner phrases may be overt, such as the determiners *the* and *a* in phrases like [DP the dog] or [DP a child], or they may be null, as in the case of plural nouns like [DP Ø dogs] or [DP Ø children]. The argument for clauses as DPs would posit that the same null determiner head that plays a role in the formation of plural DPs could also play a role in the formation of DPs from subordinate clauses. The data collected in this project provide more evidence through coordination that DPs and CPs have very similar syntactic distributions.

While NP+SBAR was also within the top ten unlike coordinations in the PTB, ADVP+PP and PP+ADVP were the most common in the PTB. Examples from the PTB are presented in (15).

(15) a. Beauregard was mentioned twice—although [ADVP very briefly] and [PP in passing].
b. A huge production system built [PP in the sea off Santa Barbara] and [ADVP ashore] is sitting idle.

ADVP+PP and PP+ADVP were within the top coordinations from the COCA data as well. Their frequent co-occurrence likely has to do with ADVP's and PP's shared purpose of adverbial modification in adjunct position. A null functional morpheme could be used to explain this coordination, and this idea would be quite similar to Bowers's Pred (predicate) proposal but applied to adjuncts of verbs instead of complements.

6.2 Differences Between Conjunct Positions

When considering each phrasal category in isolation and controlling for their different total frequencies, in both the COCA and the PTB, NPs had a very strong tendency toward being in the first conjunct position, and SBARs had a very strong tendency toward the second conjunct position. In the COCA data, VPs had a very strong tendency toward the second conjunct position, and ADJPs had a moderate tendency toward the first conjunct position.

It seems like phrasal categories that can be very short, like NPs, are more likely to appear as the first conjunct, but longer phrases, like CPs or VPs, are more likely to be the second conjunct. This may be related to a phenomenon called *heavy NP shift*, in which a noun phrase appears to the right of its expected canonical position due to its "weight" (Kayne, 1994, Chapter 7). Example (16) explores heavy NP shift through *prepositional dative constructions*, where the recipient of a ditransitive verb (in this case, "Jen") is the object of the preposition *to* (Colleman et al., 2010).

(16) a. I gave [NP the large book of poems] [PP to Jen].
b. I gave [PP to Jen] [NP the large book of poems].

All of the constituents in (16a) appear in their canonical, expected positions; the direct object noun phrase appears closest to the verb, and the prepositional phrase containing the recipient is after the NP. In (16b), heavy NP shift moves the direct

object NP into a position after the PP. Shifting can only occur if the NP is long and complex; when it is short, shifting is prohibited, as (17) shows.

(17) a. I gave [NP it] [PP to Jen].
b. * I gave [PP to Jen] [NP it].

Shifting can also target syntactic categories other than noun phrases. In (18a), the complement and adjunct of the noun "statue" appear in their expected positions, with the complement [PP of him] closer to the noun. In (18b), the complement is heavier than the adjunct and thus appears further to the right.

(18) a. the statue [PP of him] [PP in the park]
b. the statue [PP in the park] [PP of that old musician from the 19th century]

The main idea behind shifting can be applied to coordination and the trends that we observed in the results section regarding asymmetry in conjunct positions. If heavier constituents undergo shifting to appear after lighter constituents within phrases, this would explain why longer and more complex conjuncts tend to appear in the second conjunct position of coordination phrases. Example (19) shows this intuition through the like coordination of two NPs with different lengths.

(19) a. I bought [NP apples] and [NP some strange looking fruits I found in the produce aisle].
b. ? I bought [NP some strange looking fruits I found in the produce aisle] and [NP apples].

We can also observe heavier constituents appearing in the second conjunct position in unlike coordinations, as shown by example (20). Although (20b) is not ungrammatical, (20a) sounds a bit more natural.

(20) a. John is [AP healthy] and [PP in the best shape of his life].
b. John is [PP in the best shape of his life] and [AP healthy].

6.3 Limitations

One shortcoming of this paper lies in the evaluation plan: the human reviewers were not blind to the labels given to coordination phrases by the parser. With more resources, a future iteration of this project could include the creation of a small gold standard dataset of coordinations and use the more formal precision, recall, and F1 metrics to gauge the parser's accuracy in the identification of coordinations. Still, the raters' evaluations reveal the limitations of an analysis that utilizes an existing constituency-based parser on raw COCA data, which includes a size of parse errors. We acknowledge the drawbacks of such an approach and have supplemented the analysis of COCA data with data from the Penn Treebank for this purpose, which is not processed using a parser. These data sources together provide more concrete examples of the possibilities of unlike constituent coordination.

7 Conclusion

This paper approached the problem of understanding the syntax of two-termed coordination phrases through a computational corpus analysis. Previous research has not attempted a thorough analysis of coordination based on English corpora, instead relying on intuitive acceptability judgments to inform their theories. We conducted a syntactic analysis by extracting coordination phrases from the Corpus of Contemporary American English and the Penn Treebank, and we investigated the most common unlike coordinations and the syntactic categories that appeared in either of the two conjunct positions.

Some of the findings from this project have interesting implications for coordination and syntax as a whole. The high frequency of coordinations of noun phrases with subordinate clauses provides further proof that noun phrases and clauses share similar syntactic distributions and may be structurally defined as determiner phrases. The tendency for first conjuncts to be shorter constituents and second conjuncts to be longer ones might suggest that shifting occurs in coordination structures as well. One of the main takeaways from these results is that there are evident syntactic distinctions between the two conjuncts of a coordination phrase, which support theories that posit an antisymmetric account for the structure of coordination.

Acknowledgments

We would like to thank Srinivas Bangalore for his suggestions and feedback as the second reader of this paper, as well as the students of his *Introduction to Machine Translation* class, who helped complete the project's evaluation plan. We also appreciate the three anonymous reviewers' careful reading of our paper and their constructive comments.

References

Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18.

John Bowers. 1993. The syntax of predication. *Linguistic Inquiry*, 24(4):591–656.

Noam Chomsky. 1981. *Lectures On Government and Binding*. Foris Publications.

Timothy Colleman, Bernard De Clerck, and Magda Devos. 2010. Prepositional dative constructions in english and dutch: A contrastive semantic analysis. *Neuphilologische Mitteilungen*, 111(2):131–152.

Mark Davies. 2015. Corpus of Contemporary American English (COCA).

Jessica Ficler and Y. Goldberg. 2016. Coordination annotation extension in the penn tree bank. *ArXiv*, abs/1606.02529.

Janne Bondi Johannessen. 1998. *Coordination*. Oxford University Press.

Richard Kayne. 1994. *The Antisymmetry of Syntax*. MIT Press.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Terje Lohndal. 2014. Sentential subjects in english and norwegian. *Syntax and Semantics*, 15:81–113.

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

Mary L. McHugh. 2015. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Anna Prażmowska. 2015. Is unlike coordination against the law (of the coordination of likes)?

Ljiljana Progovac. 1998a. Structure for coordination: Part 1. In *GLOT International 3.7*.

Ivan Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171.

Edwin S. Williams. 1981. Transformationless grammar. In *Linguistic Inquiry 12*, pages 645–653.

Cyril Edward Zoerner. 1995. *Coordination: The Syntax of & P*. UCI dissertations in linguistics. University of California, Irvine.

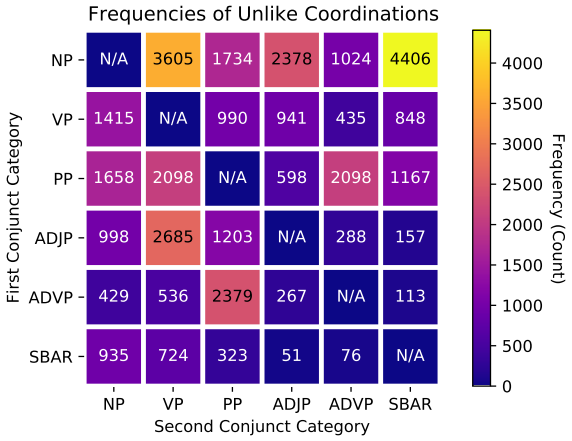


Figure 6: Heatmap displaying raw frequencies of all unlike category combinations (from COCA data).

A Heatmap of Unlike Coordinations in COCA

For completion, we include the frequency distribution of unlike coordinations for all 30 combinations of categories in the COCA data. Figure 6 visualizes these data in the form of a heatmap.

B Top Unlike Coordinations by Genre and Conjunction

The figures is in this appendix display the most frequent unlike category coordinations for each COCA genre and for each type of coordinating conjunction (*and*, *or*, *but*, *nor*) from the COCA data. Figures 7, 8, 9, 10, and 11 correspond to each of the five COCA genres, and the coordination frequencies are taken relative to all unlike coordinations within that genre. Figures 12, 13, 14, and 15 correspond to each of the four coordinating conjunctions, and the coordination frequencies are taken relative to all unlike coordinations that use the given conjunction.

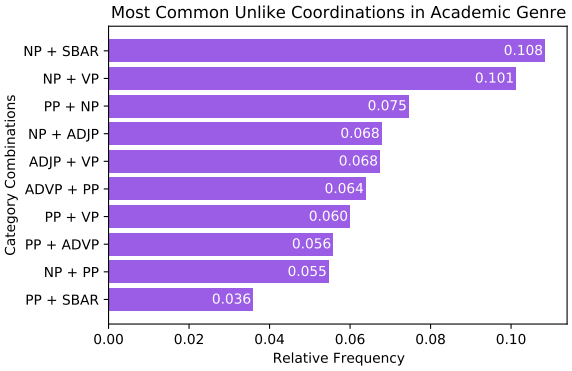


Figure 7: Academic genre.

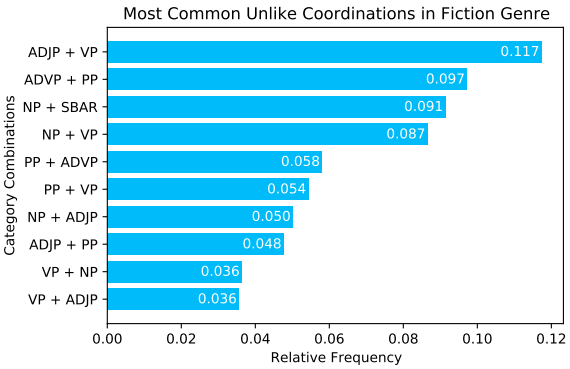


Figure 8: Fiction genre.

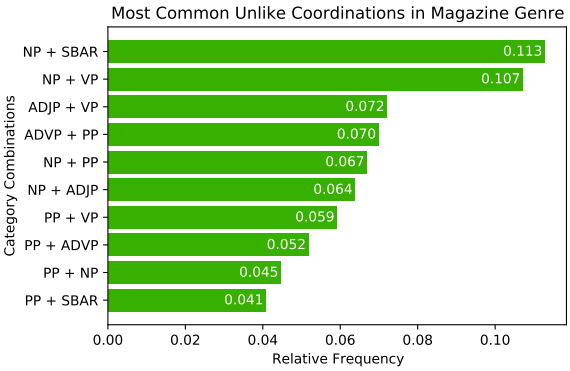


Figure 9: Magazine genre.

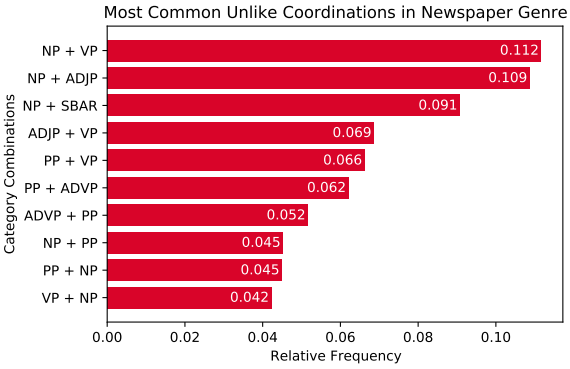


Figure 10: Newspaper genre.

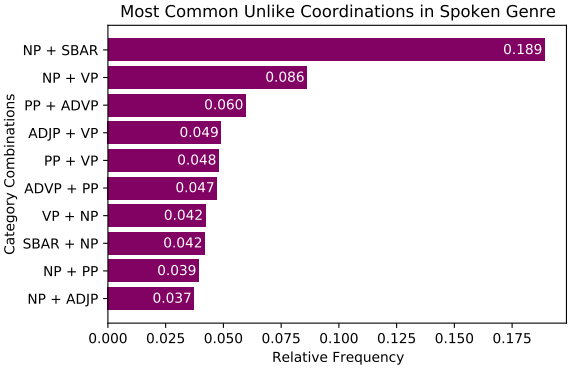


Figure 11: Spoken genre.

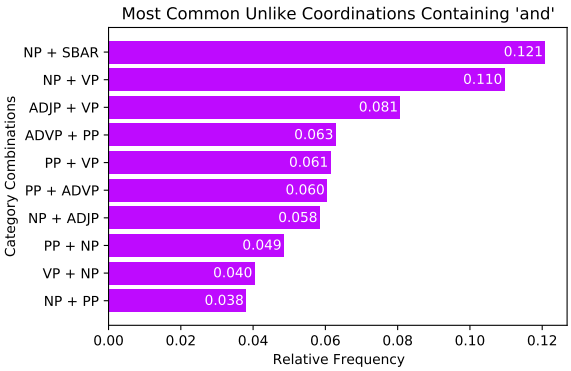


Figure 12: Unlike coordinations using *and*.

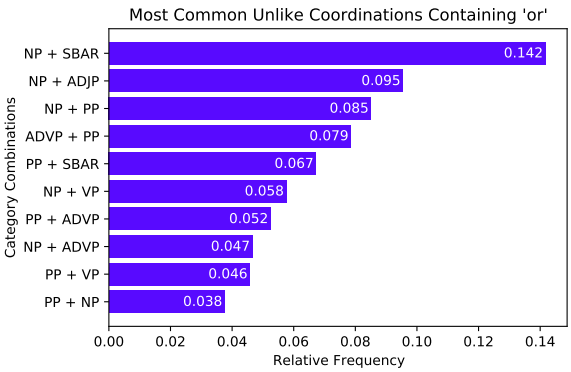


Figure 13: Unlike coordinations using *or*.

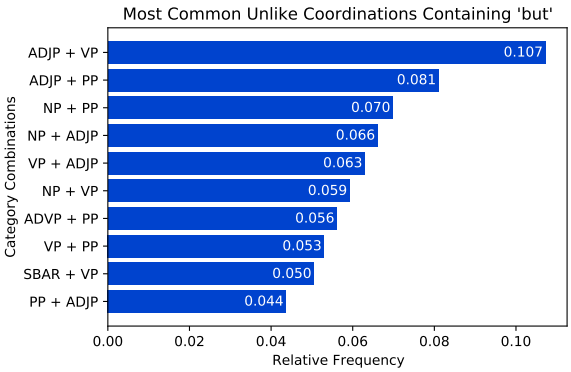


Figure 14: Unlike coordinations using *but*.

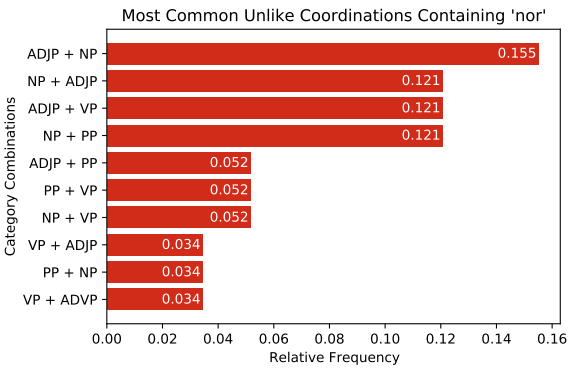


Figure 15: Unlike coordinations using *nor*.