ROLE OF PUNCTUATION IN DISAMBIGUATION OF COORDINATE COMPOUNDS*

Young-Gie Min
Department of Linguistics
University of Arizona
Tucson, AZ 85721, USA
ymin@u.arizona.edu

ABSTRACT

I discuss the role of punctuation in parsing various kinds of English coordinate compounds and examine the interaction of punctuation marks with other principles of coordinate compound disambiguation in this paper. I show that use of punctuation can greatly reduce or eliminate ambiguity in coordinate compounds in CCs, especially in unmixed CCs and short mixed CCs. To disambiguate long mixed CCs, however, one needs to consider other factors, such as semantic and pragmatic relations between members, as well as punctuation.

1. INTRODUCTION

Coordinate compounding is a syntactic device found in all natural languages. A **coordinate compound** (henceforth **CC**) consists of two or more elements of the same grammatical type or sequence of types,[1] which I call **members**. I refer to the members that occur on the leaves of any of the trees representing the structure of a CC as its **ultimate members**. Coordinate compounds are understood as **conjunctions** or as **disjunctions** of their members. Associated with particular members may be coordinating particles such as *and* and *or*, which I call **connectives**.[2] The role of connectives is to specify, among other things, whether the coordinate compounds of which they are a part are to be understood as disjunctions or as conjunctions. Two coordinate compounds are of the **same coordination type** if they are both conjunctions or both disjunctions; otherwise they are of the **opposite coordination type**. Following Langendoen (1991), a **mixed coordinate compound** is recursively defined as a coordinate compound which has a member of the opposite coordination type or which has a member which is a mixed coordinate compound. An **unmixed coordinate compound** is defined as a coordinate compound which is not a mixed coordinate compound.

CCs propose a major challenge for natural language processing as the number of possible interpretations of a CC often grows exponentially.[3] However, native speakers of a language can decide the preferred interpretation of a long CC without much difficulty, especially when punctuation marks (commas and semicolons) are used. Compare the three phrases in (1).

(1) a. Alice and Bob or Carol
b. Alice and Bob, or Carol
c. Alice, and Bob or Carol

* I thank D. Terence Langendoen for helpful comments on and discussion of this paper. All errors, of course, are exclusively the author's.

[1] I assume, following Gazdar *et al*. (1985) and Sag *et al*. (1985) that the compounding of elements of apparently different types can be analyzed as involving the compounding of elements of the same type. I do not consider coordinate compounds of sequences of elements; see van Oirsouw (1987) for discussion.

[2] Other terms for these particles are coordinators (Bloomfield, 1933:195) and conjunctions (Gazdar *et al*., 1985:170).

[3] See Langendoen (in preparation) for determining the number of possible interpretations for various types of coordinate compounds.

It is clear that (1a) is ambiguous while (1b) and (1c) are not. These simple examples illustrate the role of punctuation in disambiguating CCs.

The goal of this paper is to identify the role of punctuation marks in parsing various kinds of English CCs and examine the interaction of punctuation marks with other principles of coordinate compound disambiguation. This paper is organized as follows. In section 2, I discuss various types of CCs and define some terminology which is necessary for further discussion. I discuss the role of punctuation in disambiguating CCs in section 3. In section 4, I discuss other principles of CC disambiguation and examine the interaction of punctuation and those principles.

2. CLASSIFICATION OF COORDINATE COMPOUNDS

Following Langendoen (1991), I further classify CCs of English into three types, according to the pattern of occurrence of connectives. All three of these types are found in the next example.[4]

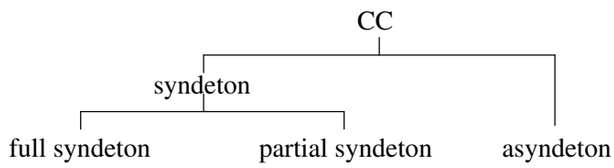
(2) Then all the chronicles and achievements of man fall humbly into perspective: all our economic competition, our strife for mates, our hunger and love and grief and war are akin to the seeking, mating, striving and suffering that hide under these fallen leaves or trees, or in the waters, or on the boughs. (APHB 324 42)

All of the coordinate compounds that are found in (2) are listed in (3).

(3) a. chronicles and achievements
b. hunger and love and grief and war
c. our economic competition, our strife for mates, our (b)
d. seeking, mating, striving and suffering
e. leaves or trees
f. under these fallen (e), or in the waters, or on the boughs

The three types of coordinate compounds in English are described as follows. First, I call a CC like (3b) and (3e) in which the same connective appears between each pair of members a **full syndeton**. Second, I call a CC like (3d) in which a connective appears before the last member only a **partial syndeton**. I refer to full and partial syndetons collectively as **syndetons**. Finally, I call compounds like (3c) with no connective between any pair of members an **asyndeton**. This classification is schematized in (4).

(4)



The example (2) exhibit the nesting of CCs within others. For example, the CCs (3b) and (3e) are contained within the larger CCs (3c) and (3f) respectively. However, the contained CCs in these examples are not themselves members of the containing CCs, but instead are proper constituents of members of the containing expressions. I refer to this type of nesting of coordinate compound structures as **submember nesting**. On the other hand, syndetons and perhaps also disjunctive asyndetons can themselves occur as a members of coordinate compounds of any type. I call this kind of nesting of coordinate com-

[4] All examples that follow are drawn from the electronic American Publishing House for the Blind (APHB) corpus unless noted otherwise. Citations refer to text and sentence number based on a lemmatization program written by Slava Katz. The coordinate compound of interest in each example is italicized.

pounds **member nesting**. The connective associated with the compound that contains nested members has **scope** over the connectives that are associated with the nested members. For example, the simplest of mixed coordinate compounds have three ultimate members and one occurrence each of the connectives *and* and *or*. These compounds have two possible interpretations, one as a conjunction with a disjunction as a nested member, and one as a disjunction with a conjunction as a nested member. In the former case, *and* has scope over *or*, and in the latter case *or* has scope over *and*.

3. ROLE OF PUNCTUATION IN DISAMBIGUATING COORDINATE COMPOUNDS

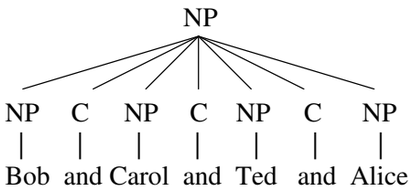
3.1 Unmixed coordinate compounds

To see how punctuation affects the interpretation of unmixed coordinate compounds, consider how the movie title in (5) is understood by a fluent reader who has not encountered it before, either as presented in (5a), or as in one of the ways in (5b) through (5h).

(5) a. Bob and Carol and Ted and Alice
b. Bob, and Carol and Ted and Alice
c. Bob and Carol, and Ted and Alice
d. Bob and Carol and Ted, and Alice
e. Bob, and Carol, and Ted and Alice
f. Bob, and Carol and Ted, and Alice
g. Bob and Carol, and Ted, and Alice
h. Bob, and Carol, and Ted, and Alice

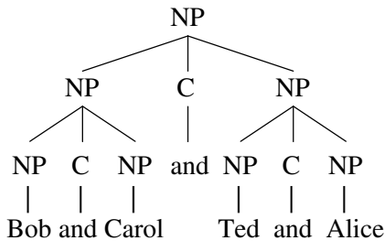
Assuming that the reader makes no strong associations between any of the names in the title, he or she will understand (5a) as having a flat structure as in (6).

(6)



If they are understood as couples then its structures will be (7).

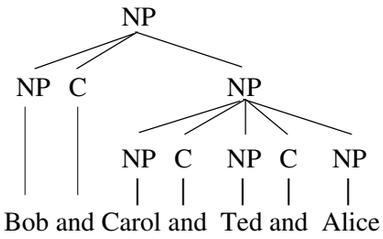
(7)



On the other hand, if one of the forms in (5b) through (5h) is presented, then he or she will understand it as having a structure that depends on the number and placement of the commas. The commas used in these examples function as what Nunberg (1990:39) calls **separator commas**, and serve to indicate constituent structure in the following way. The members that precede a comma up to the next preceding comma or the beginning of the CC, whichever comes first are grouped together; similarly the members

that follow a comma up to the next following comma or the end of the CC, whichever comes first are grouped together. For example, the structure that is associated with the CC in (5b) is (8).

(8)



If separator commas appear before each connective in a full syndeton, it indicates that no member nesting is intended. For instance, the only possible structure for (5h) is (6).

Interpretations which exhibit second degree member embedding can be distinguished by punctuation if, for example, one could use the semicolon to separate elements at the top level of embedding and the comma to separate elements at the next level of embedding. For instance, the four interpretations of (5a) which involve second degree member embedding are represented in (9a) through (9d).

(9) a. Bob; and Carol, and Ted and Alice
b. Bob; and Carol and Ted, and Alice
c. Bob, and Carol and Ted; and Alice
d. Bob and Carol, and Ted; and Alice

Such use of the semicolon and the comma together to indicate second degree member embedding, however, is almost never used, and no higher degrees of member embedding can be indicated using punctuation marks.[5]

Nunberg (1990:36) contrasts separator commas with **delimiter commas**, which are used especially in disjunctive CCs to mark off or delimit a member which is appositional to another. A comma or semicolon preceding the connective *or* may be used to indicate that the following member is to be understood as a paraphrase of the preceding member, as in (10).

(10) a. What is the *value, or rental price*? (APHB 352 569)
b. The house is big - *almost half an acre, or about ten times the size of a suburban house.* (APHB 75 1627)
c. He had been marked for death by a fellow prisoner, Vito Genovese, the most feared *capo, or boss*, of the entire Cosa Nostra. (APHB 315 16)

When the following member is not the final constituent of the sentence in which it appears, that member is followed by another comma, as in (10c). In contrast, when the comma that precedes *or* does not have this function, a comma does not necessarily follow the following member when it is nonfinal. An example in which no such comma appears is given in (11).[6]

(11) *To kill, to steal, to lie, or to covet another person's possessions* still leads to varying degrees of misery for the victim and the perpetrator. (APHB 467 463)

Finally, commas (occasionally, semicolons) precede certain connectives and not others in certain compounds to indicate member nesting as in (12).

---

[5] See Langendoen (1987) for similar remarks regarding intonation.
[6] In (11), commas separate the members of a partial syndeton. The use of punctuation for this purpose is obligatory.

(12) a. You smell the shop before you see it; it's a lovely smell, I can't articulate it easily, but it combines *[[must and dust and age], and [walls and floors of wood]]*. (APHB 215 149)
b. *[[I took the whole thing and threw it in his face], and I jumped over the counter, and [there was a fork and I used that to stab him in the face]]*. (APHB 35 1399)
c. It enhances *[[stocks and broths], and soups of [fish, poultry, meat, lentils and beans]; also [tomato and other vegetable soups]]*. (APHB 375 232)

The punctuation is used **selectively** in such compounds. Selective use of punctuation generally indicates that the unpunctuated sequences of ultimate members and connectives that immediately precede and follow a punctuated connective constitute members of a compound of which the punctuated connective is a connective. Selective use of commas thus indicates member nesting. It should be noted that since any connective occurring in one of the unpunctuated sequences is associated with a nested member of that compound, the punctuated connective has scope over the connectives that are associated with the nested members.

3.2. Mixed coordinate compounds

In this section, I analyze the structure and interpretation of mixed coordinate compounds. If punctuation marks are used selectively in CCs, then some potential ambiguities are resolved. Commas can be used selectively in mixed coordinate compounds in two partly contradictory ways: first, to indicate member nesting; and second, to indicate paraphrase relationships between members or to set off the second member of a disjunction.

3.2.1 Use of punctuation to indicate member nesting

In mixed CCs, commas can be used selectively to indicate, at least in part, the member nesting associated with its intended interpretation. I begin by considering the use of punctuation marks in mixed coordinate compounds with three ultimate members. In (13) through (16), I schematize the four ways in which a mixed coordinate structure with three ultimate members can be disambiguated by selective use of a comma to indicate member nesting. An example is provided after each schema.

(13) a. *X or Y, and Z* is understood as *[X or Y] and Z*
b. Add *fish stock or water, and salt*. (APHB 375 2037)

(14) a. *X, and Y or Z* is understood as *X and [Y or Z]*
b. *Tom knew his place, and kept it or suffered severely for breach of discipline*. (APHB 173 5307)

(15) a. *X and Y, or Z* is understood as *[X and Y] or Z*
b. The bush babies *cluster together and groom each other, or run through the trees in gangs*. (APHB 74 741).

(16) a. *X, or Y and Z* is understood as *X or [Y and Z]*
b. A club must have *a chairman, or a committee and a chairman*. (APHB 478 211)

3.2.2 Paraphrase relations in mixed coordinate compounds

In mixed CCs, punctuation can also be used selectively to indicate paraphrase relations between members. The schematic forms these compounds take and examples that illustrate it are listed in (17) through (19).

(17) a. *X, or Y, and Z* may be understood as *[X or Y] and Z*
b. Spiles, or stakes, and planks seemed to be the answer. (APHB 361 1057)

(18) a. *X, or Y and Z,* may be understood as *X or [Y and Z]*

b. In the present discussion, *the second, or broader and less mystical*, meaning is accepted. (APHB 283 1399)

(19) a. *X and Y, or Z* may be understood as *X and [Y or Z]*
b. The man was *as white as a piece of paper and trembling, or jerking*. (APHB 553 2890)

It should be noted that the member nesting of the compound schematized in (19) is the opposite of the member nesting of the compound schematized in (15), despite the identity of their punctuation patterns. Thus, this punctuation pattern provides no decisive clue as to its interpretation, and other clues are necessary in order to determine the correct member nesting.

3.2.3 Selective use of punctuation in mixed partial syndetons

Mixed partial syndetons are hard to disambiguate using punctuation alone, in part because punctuation must be used whenever no connective appears between members (See note 6). Compare the examples in (20) and (21).

(20) a. Was there *[a dusty silence or [a roaring storm, tossing trees and driving rain]]*? (APHB 116 1658)
b. Add the *[chicken meat and [cooked rice, noodles or vermicelli]]*. (APHB 375 1690)

(21) a. Add *[[broth or bouillon], wine and thyme]*. (APHB 375 2110)
b. In the evenings we enjoyed *[[dinner and drinks], dancing or a movie]*. (APHB 328 5558)

The examples (20a) and (20b) show the same punctuation pattern as the examples (21a) and (21b) respectively, but the structures associated with them are different. More such examples are in (22).

(22) a. Combine with *[sugar, orange juice and [champagne or white wine]]*.
b. It includes *[baked potato, [coffee or tea] and ice cream]*. (APHB 19 2379)

In short, punctuation alone cannot disambiguate many mixed partial syndetons, and other principles are needed to handle them.

4. INTERACTION WITH OTHER DISAMBIGUATION STRATEGIES

4.1 Syntactic strategy

Plural context may disambiguate mixed CCs. For example, the preposition *between* governs a plural noun phrase or a conjunction of noun phrases. Thus a mixed CC of noun phrases governed by *between* cannot be analyzed as a disjunction if at least one of its members is singular.

(23) The drop varied between *1/2 and 1/4 or 1/8*. (APHB 136 1092)

Other plural contexts are those provided by the adverb *together* and certain uses of verbs like *combine* and *mix*.

Other syntactic contexts that can disambiguate mixed CCs are those in which a disjunctive choice is required. For example, the particle *either* takes a disjunction as its complement, and as a consequence, the compound in (24) is syntactically disambiguated.

(24) The residents she had either *known and discarded or never been interested in*. (APHB 264 5598)

A context similar to the one provided by *either* is provided by the complementizer *whether*.

Syntactic properties of the ultimate members sometimes give clues as to the intended interpretation of mixed coordinate compounds, as in the next examples.

(25) a. You'd have sold him out to *me or [Royce and Benson]*. (APHB 385 4242)
b. Would they want me to call them *[aunt and uncle], or by their first names*? (APHB 281 398)

The presumed intended interpretation of (25a) is as a disjunction. Under this interpretation, the two proper names are grouped together as members of a nested member of the disjunction, whose other member is a personal pronoun. This arrangement is not required by the syntax of English, but is suggested by the lexical categorizations of the ultimate members. Similarly, the intended interpretation of (25b) as a disjunction is indicated not only by the punctuation, but also by the fact that the members of the nested members are of a different grammatical category from that of the final member.

However, such syntactic clues are not decisive. The examples in (26) cannot be disambiguated based on syntactic clues alone.

(26) a. As always, it is important to distinguish between *tasters or experimenters and chronic inhalant abusers*. (APHB 435 268)
b. Besides most of these occult documents were either *pagan or heretical and therefore thoroughly evil*. (APHB 452 133)
c. Sybil never found him *alone and quiet or in a mood to talk*. (APHB 381 3796)

4.2 Semantic and pragmatic strategy

In many cases, punctuation and syntac information do not provide enough clues, and one need to consider semantic and pragmatic information to disambiguate CCs. Langendoen and Min (1995) proposed that the preferred interpretation of CCs can be determined from the following principles.

(27) a. **Natural class disjunction**: Prefer an interpretation of a mixed CC in which the members of a disjunction form a natural class.
b. **Natural whole conjunction**: Prefer an interpretation of a mixed CC in which the members of a conjunction form a natural whole.

Consider the natural interpretation of the mixed CC in (28).

(28) The man must be *rich or young and good-looking*. (from W. Hodges, *Logic*, 1977, p.78)

(29) a. [rich or young] and good-looking.
b. rich or [young and good-looking].

Since *young* and *good-looking* form a natural whole to a greater degree than *rich* and *young* form a natural class, (27b) predicts that (29b) is the preferred interpretation of (28).[7]

As I pointed out in section 3.2.2, some punctuation patterns provide no decisive clue as to the interpretation of a mixed CC. The two principles may be used to disambiguate those cases. Consider (19b) again.

[7] The lexicon in this model is a directed acyclic graph (DAG) with one root node. Given three nodes A, B, and C, we may know which two of them are more closely related by comparing the *lowest common ancestor* (LCA) of A and B with the LCA of B and C in the DAG. If the LCA of A and B is lower than that of B and C in the DAG, then A and B are more closely related than B and C.

(19) b. The man was *as white as a piece of paper and trembling, or jerking*. (APHB 553 2890)

According to (15a) and (19a), the CC in (19b) can be interpreted either as (30a) or as (30b).

(30) a. [as white as a piece of paper] and trembling, or jerking
b. as white as a piece of paper and [trembling, or jerking]

In (19b), *trembling* and *jerking* form a natural class to a greater degree than *as white as a piece of paper* and *trembling* form a natural whole. As a result (30b) is the preferred interpretation of (30b). Another example of this sort is (31), in which a semicolon is used instead of a comma.

(31) *Chill and serve with a plain mayonnaise to which you have added fresh dill to taste; or serve with Green Mayonnaise.*

Applying (15a) and (19a) result in two different interpretations, but the principle (27a) correctly selects the interpretation in which *or* has scope over *and*.

5. CONCLUSION

In this paper, I discussed the role of punctuation in parsing English coordinate compounds. Use of punctuation can greatly reduce or eliminate ambiguity in CCs, especially in unmixed CCs and short mixed CCs. On the other hand, many mixed partial syndetons cannot be easily disambiguated based on punctuation patterns and one needs to consider other factors such as syntactic, semantic and pragmatic relations between members. The two principles based on the semantic relation between neighboring members are very useful when there is no other clue or when other disambiguation strategies give conflicting results.

REFERENCES

Bloomfield, Leonard. 1933. Language. Henry Holt. New York.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar.* Basil Blackwell. Oxford.

Langendoen D. Terence. 1987. On the Phrasing of Coordinate Compound Structures. In Brian Joseph and Arnold Zwicky (*eds). A Festschrift for Ilse Lehiste.* 1987. pp. 186-196. Ohio State University.

Langendoen D. Terence. 1991. An Analysis of Coordinate Compound Expressions in English. unpublished manuscript. University of Arizona

Langendoen (in preparation). Structure of Coordinate Compounds.

Langendoen D. Terence and Young-Gie Min. 1995. Disambiguation of Coordinate Compound Expressions. Paper presented at the Eighth CUNY Sentence Processing Conference. 1995.

Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation.* CSLI Lecture Notes 18, Stanford, California.

Sag, Ivan A., Gerald Gazdar, Thomas Wasow and Steven Wechsler. 1985. Coordination and How to Distinguish Categories. *Natural Language and Linguistic Theory* 3, pp. 117-171.

van Oirsouw, Robert R. 1987. *The Syntax of Coordination.* Croom Helm. London.