COORDINATION: SOME PROBLEMS AND SOLUTIONS FOR THE ANALYSIS OF ENGLISH WITH AN ATN

Lee Ann Schwartz

Pan American Health Organization
525 23rd St., N.W. Washington, D.C. 20037

Abstract

Coordination is a complex phenomenon that poses many problems for the parsing of English by computer. This paper examines some of these problems and suggests solutions within the framework of ATN parsing. Examples of complex coordination phenomena, extracted from texts translated by ENGSPAN™, the Pan American Health Organization's English-Spanish machine translation system, are presented. Schemata for extending simple networks to accommodate coordinate constructions are presented, and strategies for parsing these constructions are discussed. Focus is centered on the complications involved in parsing constructions with more than two conjuncts.

1. Introduction

The coordinate construction is one of the most common constructions in the English language. At the same time, it is one of the most complex. The parse of a sentence, i.e. the application of the rules of a grammar to that sentence in such a way as to obtain a description of it (Dowty 1985), is complicated by those aspects of coordination that have enabled it to escape formalization.

This paper examines several of the complexities that coordination introduces into parsing. *It* shows how a simple ATN grammar can be modified, and how parsing strategies can be implemented, so as to make possible the analysis of sentences exemplifying a variety of coordination phenomena.

The examples provided in the text have been extracted, whenever possible, from texts translated by the Pan American Health Organization's English-Spanish machine translation system (ENGSPAN™). The discussion of parser design and implementation of parsing strategies has as its foundation the design and parsing strategies of ENGSPAN's ATN parser.

2. Basic Characteristics of Coordination

Definitions of coordination are almost as numerous as grammars of English. For the purposes of this discussion, a coordinate construction will be defined as the product of the linking of linguistic units of similar type by a conjoining comma and/or a coordinate conjunction. A coordinate conjunction, for its part, will be defined as an element of the following set: [and, or, but] (with the understanding that 'but' is not a coordinate conjunction when it has the sense of 'except', as in "all but one").

The above characterization of coordinate constructions is purposely vague. How are the conjuncts of a coordinate construction similar? Commonly, the answer given to this question is that they are similar in syntactic type—clauses are conjoined with clauses, noun phrases with noun phrases, verbs with verbs, etc. As the following examples illustrate, the conjuncts of a coordinate construction do tend to be of the same syntactic type. (In these examples the coordinate constructions are delimited by parentheses and are explicitly marked for type.)

(1) They stimulate close co-operation between (NP the Ministries of Health, other health institutions, non-governmental organizations, the civil defense and the representatives of the international community) (preps both before and during) emergency situations caused by (adj natural or man-made) disasters.

(2) These recommendations emphasize (CL that the program strategy remains unchanged and that to achieve program goals it is necessary to (VP maintain high levels of

vaccination coverage, implement ($_{NP}$ intensive surveillance and active case investigation), and institute aggressive outbreak control)).

The coordinate constructions found in these sentences are examples of the simplest type of coordinate construction—the constituent coordinate construction (Stockwell, Schachter, and Partee 1973). From a surface structure perspective, the conjuncts of these constructions are single constituents that are of the same syntactic type and are immediately adjacent either to a coordinate conjunction or to a conjoining comma. This paper will focus on the analysis of constructions of this type, which illustrate several of the complexities that coordination introduces into parsing, including the following:

1. Coordination can take place at any syntactic level.

2. There is no limit to the number of coordinate constructions that can occur in a single sentence.

3. Coordinate constructions can have any number of conjuncts.

4. Coordinate constructions are often embedded within other coordinate constructions.

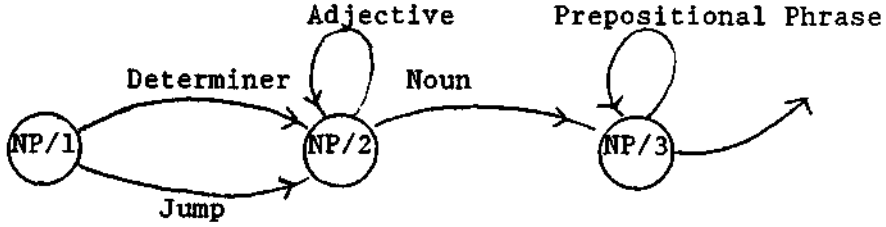
5. The beginning and end of coordinate constructions are commonly unmarked.

3. Equipping an ATN for the Analysis of Constituent Coordinate Constructions

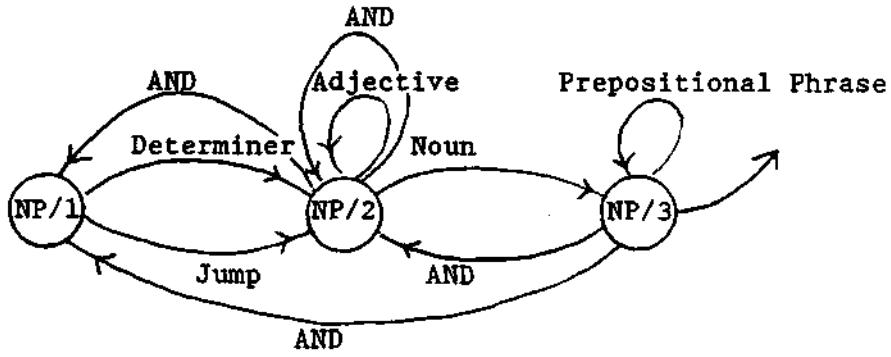
A parser must be prepared to find a coordinate conjunction, and it must be prepared to begin the parse of a coordinate construction, at virtually any point in the analysis of a sentence. There are two basic ways in which the network system of an ATN that is equipped only for the analysis of simple and complex constructions can be modified for the analysis of constituent coordinate constructions: brute-force extension and final-state extension.

The brute-force extension of an ATN, as described by Boguraev (1983), is accomplished with the addition of conjunction arcs to virtually every state of its networks. The simple NP network of Figure 1, for example, is extended by brute force to produce the compound network of Figure 2. Equipped with this compound network, an ATN parser can analyze both simple NPs and constructions in which complete NPs, or any parts thereof, are conjoined.

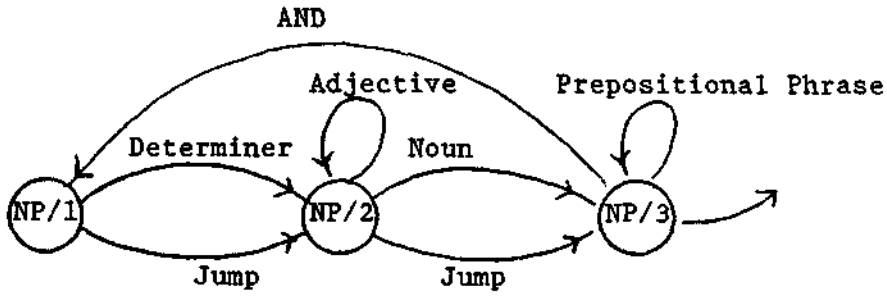
The final-state extension of an ATN, as described by Blackwell (1981), is accomplished with the addition of AND word arcs to the final states of each network. The final-state extension of the simple



Simple NP Network
Figure 1



Brute Force Extension
Figure 2



Final-State Extension
Figure 3

NP network of Figure 1 in accordance with Blackwell's scheme would result in the compound NP network of Figure 3.

Final-state extension equips an ATN for the analysis of compound constructions in which the conjuncts are complete phrases that correspond in type to one of the networks. For the analysis of constructions in which constituents below the phrase level are conjoined, internal-state modification must accompany final-state extension. Internal-state modification adds jump arcs to a network in such a way as to enable the parser to reach the final state of that network without having parsed elements that would be essential for the successful analysis of a simple phrase. Internal-state modification is responsible for the presence of the jump arc that originates at NP/2 and terminates at NP/3 of Figure 3. This arc makes it possible for the parser to analyze a conjunct without a head noun (as it must do, for example, in the analysis of conjoined adjectives).

Just how many jump arcs must be added to a network and what conditions must be associated with their traversal will differ from network to network. Whatever the network, however, these arcs will be few in number. When the parser analyzes a conjunction, it can make substantial progress towards reaching the final state of the network, where it finds the conjunction arc, by traversing the jump arcs that it traverses in the analysis of complete simple phrases.

Brute-force extension and final-state extension, as just described, only equip an ATN for the analysis of sentences with constituent coordinate constructions consisting of two conjuncts joined by 'and'. The ability of parsers that have been modified in these ways to analyze natural English text is quite limited. To even have a chance at successfully parsing English text, a parser must be able to analyze coordinate constructions with more than two conjuncts, and it must be able to analyze constructions marked by correlative conjunctions ('both', 'either', 'neither').

An ATN parser can be equipped for the analysis of coordinate constructions in which the conjuncts are linked by correlative conjunctions with the addition of correlative conjunction arcs to its networks. These arcs would have to be added to virtually every state of an ATN that is extended by brute force. Such an ATN, already burdened with a multitude of conjunction arcs, would be overburdened with these arcs. In fact, with each additional complexity of coordination that the grammar was modified to account for, it would would become more and more unmanageable in size. At this point, then, brute-force extension is excluded as a viable solution to the problem of equipping an ATN for the analysis of compound constructions. Attention is focused on final-state extension.

An ATN that has undergone final-state extension and internal-state modification can be equipped for the analysis of constructions marked by correlative conjunctions with the addition of a correlative conjunction arc to the initial state of each of its

networks. This modification to the network system greatly enhances the parser's ability to analyze coordinate constructions without overburdening the networks with arcs. The parser can be equipped for the analysis of constructions in which the conjuncts are linked by coordinate conjunctions other than 'and' with a simple replacement of the AND word arc with a coordinate conjunction category arc. The modifications that would have to be made to equip the parser for the analysis of constructions with three or more conjuncts are not, however, so simple.

Final-state extension, as described above, is employed by Blackwell for the analysis of coordinate constructions with two and only two conjuncts. The Blackwell parser performs an iterative analysis of coordinate constructions. As it analyzes conjuncts, the parser moves from the initial state of a network to the final state and back again to the initial state without traversing either a send arc or a seek arc. This, Blackwell notes, complicates procedures for creating and manipulating registers.

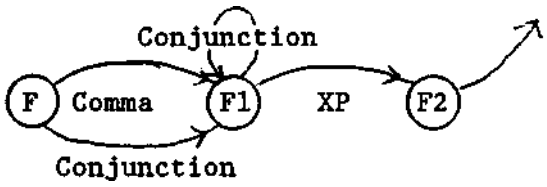
A way to keep from complicating register creation and manipulation and to not have to limit the input to two-conjunct constructions is to implement a recursive analysis. This can be done by replacing the conjunction arc that loops from the final state of a network back to the initial state with a conjunction arc that originates at the final state and terminates at a state from which a conjunct seek arc originates.

The step up from parsing a construction with two conjuncts to parsing a construction with any number of conjuncts introduces a major complication into the parsing process that is not related to the creation and manipulation of registers: how to deal with commas. Unlike coordinate conjunctions, the comma has several possible functions. In addition to separating the conjuncts of some coordinate constructions with two conjuncts and all coordinate constructions with more than two conjuncts, commas separate adverbials from a preceding and/or following string; they delimit restrictive relative clauses; and they signal apposition.

The function of a comma in a sentence can only be determined by parsing. The extension of a grammar to include comma arcs provides for a more efficient and more accurate analysis than its alternative, the stripping of commas from the input before analysis. A parser that strips commas from its input makes the already difficult task of determining the boundaries of constituents in an English sentence even more difficult. It artificially increases the potential for ambiguity in the language. The addition of comma arcs to a network grammar increases the size of that grammar, but with the strategic placement of these arcs and definition of networks the increase can be minimal.

There are several possible schemes for final-state extension in which conjoining comma arcs, conjunction arcs, conjunct seek arcs, and compound send arcs are added to the final states of ATN networks. In

the following scheme, state F represents the simple send state, i.e. the final state of the network before extension. The other F states are added during extension.



The XP arc is the conjunct seek arc. For the analysis of constructions in which the conjuncts are of identical syntactic type, this arc will be of the same type as the network to which it is added. On the NP network it will be an NP seek arc, on the VP network it will be a VP seek, etc. The comma and conjunction arcs in the scheme need no explanation. The compound send arc is the second send arc in a network (the first being the simple send arc). It enables the parser to conclude the analysis of a coordinate construction, and it makes it possible for the parser to place different conditions on the conclusions of the analyses of coordinate and simple constructions, properly constraining each.

An ATN that has undergone final-state extension in accordance with the scheme illustrated above and has also undergone initial- and internal-state modification is equipped to parse a sentence with any number of constituent coordinate constructions (marked or not marked by correlative conjunctions), with any number of conjuncts. These conjuncts may correspond in type to any network or category arc (except the comma and conjunction category arcs) in the network system.

4. Strategies for Parsing Coordinate Constructions

Once an ATN is equipped with the arcs it needs to parse coordinate constructions, it must have a strategy to follow in traversing those arcs. The strategy followed by a serial ATN parser is determined by the scheduling of its arcs. The strategy for parsing coordinate constructions in particular is determined by the relative order in which the send arc and the comma and conjunction arcs originate from the simple send states of the networks. This order determines whether the parser will attempt to conclude the parse of the construction it is analyzing when it encounters a coordinate conjunction or a potential conjoining comma, or whether it will continue the analysis of that construction, attempting to parse a sister conjunct for it. The scheduling of arcs from the simple send states has a noticeable effect on the efficiency with which sentences of the following types are parsed:

(3) The council has reviewed the progress of the program and the two groups that advise the Organization on this issue have met five times.

(4) These anniversaries and events provide another opportunity to review and promote WHO and PAHO program goals and to strengthen national mobilization for health development.

In an ATN analysis of (3), after the parser has analyzed "the program" on the NP network, the question is whether it should traverse the conjunction arc on that network and attempt to parse the post-conjunction string as a sister conjunct of that phrase, or whether it should traverse simple send arcs to conclude the analysis of "the council has reviewed the progress of the program" and attempt to analyze the post-conjunction string as a clausal conjunct. The first course of action would lead the parse to block after "the two groups that advise the Organization on this issue" had been analyzed as part of a conjoined prepositional object. The second course of action, resulting from the ordering of simple send arcs before conjoining comma and coordinate conjunction arcs, would lead to a more efficient parse of this sentence.

The scheduling of simple send arcs before conjoining comma and conjunction arcs leads to a parsing strategy in which post-conjunction strings are analyzed first at the highest levels of recursion possible, and then, if necessary, at lower levels of recursion. In the simplest of coordinate constructions, however, coordination takes place at the lowest level of recursion, that is, coordination joins the words immediately to the right and left of the conjunction. The parse of (4) will only proceed smoothly if conjunction arcs are traversed before simple send arcs, not after. The opposite ordering of arcs will lead not only to a substantial amount of backtracking, but to several false starts, i.e. several attempts by the parser to analyze a construction of a certain type without having evidence of the existence of a construction of that type.

In general, the ordering of send arcs after the conjoining comma and conjunction arcs originating from the same state will result in the most efficient analysis of a text. And although it will not result in the most efficient analysis of (3), at least it will not lead the parser to pursue analyses without reason. The post-conjunction string in (3) does begin with an NP. The parser may misanalyze that NP as a conjunct of a conjoined prepositional object. When it discovers its error, however, by placing the NP on a well-formed phrase list, the parser can incorporate it into the analysis without reparsing it. After the parser posits the existence of a clausal conjunct, it will retrieve the NP from the well-formed phrase list (as the subject of the clause) and quickly proceed to a successful conclusion of the analysis of the sentence.

The same choices that are available to a parser when it encounters a coordinate conjunction in a sentence are available to it when it encounters a comma. The relative ordering of conjoining comma and send arcs from the simple send states of networks should therefore be the same as the relative ordering of conjunction and send arcs from

those states. Because the possibility exists, however, that the comma does not function as a conjoining comma, there are even more options that the parser must choose from in determining how and where in the network to analyze the comma. Attempts to parse the commas in the following sentence as conjoining commas, at whatever level of recursion, would result in failure.

(5) As we have heard in past presentations, over the past ten years, the area of health and behavior has become an integral part of the activities.

Misanalyses of non-conjoining commas can be reduced if the condition on the traversal of conjoining comma arcs requires either that a coordinate conjunction appear somewhere after the comma in the sentence or that the elements immediately to the right and left of the comma be potential adjectival or adverbial modifiers (which can be conjoined in series without a coordinate conjunction). Such a condition would increase the efficiency of a parse of a sentence that included commas, but no coordinate constructions (at the expense of only a handful of stylized coordinate constructions in which parallelism obviates the need for an explicit coordinate conjunction). It would not, however, increase the efficiency of the analysis of (5). Additional conditions might be placed on the traversal of the arcs of the network system in order to enforce a strategy whereby, if the parser is analyzing an adverbial or a relative clause, it attempts to conclude the analysis of that construction before it tries to traverse a conjoining comma arc. Such a strategy would work well for those sentences in which the adverbial or relative clause was not itself, or did not itself include, a coordinate construction. It would not work well otherwise. Experience has suggested the advisability of implementing the strategy of concluding the analysis of adverbials and relative clauses as soon as it is possible to do so.

Just as a coordinate conjunction and a comma introduce ambiguities into a sentence, so does a potential correlative conjunction. The words 'both', 'either', and 'neither' can all function not only as correlative conjunctions, but also as pronouns or determiners; 'either' can function as an adverb as well. A condition on the traversal of correlative conjunction arcs requiring that the proper conjunction ('and' for 'both', 'or' for 'either', and 'nor' for 'neither') appear somewhere after the potential correlative in the sentence will help prevent misanalyses of these words. A correct and efficient analysis of sentences with potential correlatives will also depend on the placement of correlative conjunction arcs in the network system and the formulation of the condition on traversal of the simple send arcs from networks. Correlative conjunctions often appear at the beginning of a coordinate construction, but not always. It would be helpful in disambiguating potential correlatives to make the conclusion of the analysis of a simple phrase contingent upon no correlative conjunction having been parsed in that phrase. This, however, would only enable the parser to analyze the second of the following sentences:

(6) Program activities will either resume before the beginning of the next session or after its conclusion.

(7) Program activities will resume either before the beginning of the next session or after its conclusion.

There is a tradeoff between efficient and comprehensive parsing. Conditions can be written and arcs can be added to a network system to allow for the analysis of all possible variations of a construction. On the other hand, a parser might be allowed sacrifice the ability to analyze some variations of constructions in order to be able to analyze others more efficiently.

5. Unexplored Territories

An ATN that has undergone final-state extension and initial- and internal-state modification is equipped to handle many of the problems that coordination poses for computer parsing. There are several varieties of coordinate constructions, however, that an ATN modified in this way could not analyze. These include coordinate constructions characterized by various types of ellipsis and constructions in which the conjuncts are of different syntactic types.

Non-constituent coordinate constructions are elliptical. They may be characterized by Gapping, Pseudo-Gapping, Raising, or Verb Phrase Deletion, as are the constructions underlined in the following examples.

(8) While many countries have made some progress in improving disease surveillance, immunization coverage remains low and drop-out rates high. (Gapping)

(9) This is necessary so that appropriate clinical histories can be taken, laboratory specimens collected, and diagnostic tests obtained on all cases. (Pseudo-Gapping)

(10) This is intended to draw attention to and create general public interest in the celebration. (Raising)

(11) Some countries have made progress in improving disease surveillance, and others have not. (Verb Phrase Deletion)

The analysis of these sentences requires, at the least, that additional seek arcs, send arcs, and possibly jump arcs be added to the extended network system, and that initialization procedures be established to recover deletion.

The analysis of other types of elliptical constructions, which defy classification, are probably best left outside the realm of computer parsing for purposes of machine translation. These

constructions can be attributed, in large part, to the television industry and to linguists. Two are given below.

(12) (It) shaves as close as a blade, or your money back.

(13) They left, and fast.

Constructions in which the conjuncts are not of the same syntactic type must be kept within the realm of constructions that a parser should be given a chance to analyze. Examples of such constructions are given below.

(14) But it is important to note the level of commitment that the governments and the external agencies have displayed. and that this constitutes a reassurance that the goals of the program and the eradication of polio from this hemisphere could be reached by 1990.

(15) These measures should be adopted by those countries classified as infected by polio or at high risk.

(16) It also serves to keep the Director appraised of key health and behavior development in science and how such developments might impact on future research needs.

The way in which final-state extension can be modified to permit the analysis of these constructions is beyond the scope of this paper.

As a final note, it must be pointed out that even within the limited realm of constituent coordination, the scheme of final-state extension and initial- and internal-state modification is not the solution to all the problems posed by coordination. This scheme is based entirely on syntax. Semantics, pragmatics, and world knowledge often provide the information that is needed to produce a correct analysis of a construction. Examples (17) and (18) can be analyzed by an ATN that has undergone final-state extension and initial- and internal-state modification. They are likely, however, to be misanalyzed. The correct boundaries of the coordinate constructions in these sentences are marked with parentheses. Possible misanalyses are indicated by underlining.

(17) Major activities being implemented at country and regional levels are related to (the acceleration of immunization programs and the strengthening of surveillance systems) for (prompt detection of suspected cases of poliomyelitis, case investigation and immediate institution of control measures).

(18) Every effort should be made to ensure their help to (strengthen the entire program and lead to development of permanent, ongoing immunization).

The misanalyses of (17) and (18) would result from conjunctions and conjoining commas being parsed at too low a level of recursion. The misanalysis of (17) would be transparent to a translation (at least from English into Spanish). That of (18) would not ('lead' would be translated as a noun rather than as a verb).

This paper has scanned the tip of the iceberg of problems that coordination poses for computer parsing. It has suggested solutions for several of these, including how to handle constructions with three or more conjuncts. The ENGSPAN™ ATN goes far in dealing with coordination phenomena, but still much work, especially in the realm of semantics, pragmatics, and world knowledge, has yet to be done.

References

Blackwell, S.A. 1981. Processing conjunctions in an ATN parser. Thesis for Master of Philosophy, University of Cambridge.

Boguraev, B.K. 1983. Recognising conjunctions within the ATN framework. In: Automatic Natural Language Parsing, ed. by K. Sparck-Jones, and Y. Wilks. Ellis Horwood.

Dowty, D., L. Karttunen, and A. Zwicky (eds). 1985. Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives. New York: Cambridge University Press.

Leon, M. and L. Schwartz. 1986. Integrated Development of English-Spanish Machine-Translation: From Pilot to Full Operational Capability. Technical Report to the U.S. Agency for International Development (Grant DPE-5543-G-SS-3048-00).

Schwartz, L.A. 1987. Coordination: An ATN Perspective. Unpublished Ph.D. dissertation. Georgetown University. Washington, D.C.

Stockwell, R.P., P. Schachter, and B.H. Partee. 1973. The Major Syntactic Structures of English. New York: Holt, Rinehart, and Winston.