

## Subdeletion in Verb Phrase Ellipsis

**Paul G. Donecker**  
Villanova University  
800 Lancaster Avenue  
Villanova, PA 19085  
donecker@monet.vill.edu

### Abstract

This paper stems from an ongoing research project<sup>1</sup> on verb phrase ellipsis. The project's goals are to implement a verb phrase ellipsis resolution algorithm, automatically test the algorithm on corpus data, then automatically evaluate the algorithm against human-generated answers. The paper will establish the current status of the algorithm based on this automatic evaluation, categorizing current problem situations. An algorithm to handle one of these problems, the case of subdeletion, will be described and evaluated. The algorithm attempts to detect and solve subdeletion by locating adjuncts of similar types in a verb phrase ellipsis and corresponding antecedent.

### 1. Introduction

A verb phrase ellipsis (VPE) exists when a sentence has an auxiliary verb but no verb phrase (VP). For example, in the sentence "Gather ye rosebuds while ye may," "may" is the beginning of a VPE. Its antecedent is "gather ye rosebuds." The research described in this paper is part of a project to automate the resolution of VPE occurrences, and also to automate the evaluation of the success of the VPE resolution (Hardt 1995).

Based on these evaluations of the algorithm, several distinct categories of error situations have been determined. We have focused on errors in which the program selects the correct head verb as antecedent. These cases can be divided into the following categories: 1) too much material included from the antecedent, 2) not enough much material included from the antecedent, 3) discontinuous antecedents, and 4) miscellaneous.

For a subset of case 1, subdeletion, an algorithm derived from (Lappin and McCord, 1990) is evaluated

in regard to the Brown Corpus.

### 2. Background

Previous studies on evaluating discourse processing (e.g., Walker, 1989; Hobbs, 1978) have involved subjectively examining test cases to determine correctness. With the development of resources such as the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993), it has become possible to automate empirical tests of discourse processing systems to obtain a more objective measure of their success. Towards this end, an algorithm was implemented in a Common Lisp program called VPEAL (Verb Phrase Ellipsis Antecedent Locator) (Hardt, 1995), drawing on the Penn Treebank as input. The portion of the Penn Treebank examined--the Brown Corpus, about a million words--contains about 400 VPEs.

Furthermore, to automatically evaluate the algorithm, utilities were developed to automatically test the output of VPEAL for correctness. The most recent version of VPEAL contained 18 sub-parts for ranking and choosing antecedents. Testing the program's performance involved finding the percentage of correct antecedents found by any or all of these algorithms. This was achieved by having human coders read plain text versions of the parsed passages, marking what they felt to be the antecedent. Antecedents selected by VPEAL were considered correct if they matched the antecedents selected by the coders.

The remainder of this paper will describe the categories of errors observed, then describe an approach to reducing one category of errors.

### 3. Categories of Errors

The most recent version of VPEAL correctly selects 257 out of 380 antecedents from the Brown Corpus. We have divided the categories into the following categories:

A. Incorrect verb: 90 cases. In these cases, VPEAL selected an incorrect head verb for the

---

<sup>1</sup> This research was supported in part by NSF Career Grant, no. IRI-9502257.

antecedent. The causes of these errors are being evaluated.

B. Incorrect antecedent but correct verb: 33 cases. VPEAL selected the correct verb to head the antecedent, but the selected antecedent was either incomplete or included incorrect information. These cases can be further divided into: 1) too much material included from the antecedent, 2) not enough much material included from the antecedent, 3) discontinuous antecedents, and 4) miscellaneous. These subcategories are described below.

1. Too much material is included from the antecedent: 11 cases.

Example (excerpt from Penn Treebank):

produce humorous effects in his novels and tales  
as they did in the writing of Longstreet and  
Hooper and Harris

VPE: did

VPEAL's antecedent: produce humorous effects in his  
novels and tales

Coder's antecedent: produce humorous effects

Normally, an entire verb phrase is selected as the antecedent. In these cases, though, part of the selected antecedent was not required by the VPE. The most common situation (6 cases), as in the above example, was subdeletion--when the VPE structure contains a noun phrase or prepositional phrase which substitutes for a corresponding structure in the antecedent verb phrase.

2. Not enough material is included from the antecedent: 10 cases.

Example (excerpt from Penn Treebank):

But even if we can not see the repulsive  
characteristics in this new image of America,  
foreigners can

VPE: can

VPEAL's antecedent: see the repulsive characteristics

Coder's antecedent: see the repulsive characteristics in  
this new image of America

By default, only text contained by the selected verb phrase is included in the antecedent. In these cases, however, human coders have selected text that is adjacent to but not parsed as contained by the verb phrase as part of the antecedent. It can be argued that these errors are not the fault of the VPEAL algorithm--that if text is parsed as not being a part of

the verb phrase then it should still not be included when the verb phrase is chosen as the antecedent. If the above prepositional phrase "in this new image of America" were parsed as part of the verb phrase-- as indeed it should have been--then the algorithm would have derived the correct antecedent.

3. Discontinuous antecedents--the correct antecedent is split into two parts: 5 cases.

Example (excerpt from Penn Treebank):

representing as I do today my wife

VPE: do

VPEAL's antecedent: representing

Coder's antecedent: representing my wife

This situation is similar to B2 in that the antecedent is incorrect because text not contained by the selected verb phrase should be included in the antecedent. In these cases, however, the reason the omitted text is not contained by the antecedent verb phrase is that an interposing phrase (in the example above, the VPE itself) occurs in the middle of the antecedent.

4. Miscellaneous: 7 cases.

#### 4. Improving Performance in the Case of Subdeletion

In this section an algorithm is described to reduce the errors in error category B1 caused by subdeletion. Subdeletion is probably the most straightforward of the error categories. The problem category occurred when prepositional phrases and noun phrases in the antecedent verb phrases were unnecessary because of analogous phrases adjacent to the VPE. The proposed solution was to check whether the VPE has a sister node that is a prepositional phrase or noun phrase. If it does, and a phrase of the same type exists as a sister node to the head verb in the antecedent, then the phrase in the antecedent is removed. This is essentially the strategy outlined by Lappin and McCord (1990). Following are the specific steps to implementing the algorithm:

1. Check if there are any prepositional phrases or noun phrases that are sister nodes to the antecedent head verb.

2. Check if there are any prepositional phrases or noun phrases that are sister nodes to the VPE head verb.

3. If a prepositional phrase or noun phrase is found in step 1, and a phrase of the same type is found

in step 2, then remove the phrase found in step 1 from the antecedent.

For example, refer to the example from error case B.1. Step 1 would locate the noun phrase "humorous effects" and the prepositional phrase "in his novels and tales" as sister nodes to the antecedent head verb "produce."

Step 2 would locate the prepositional phrase "in the writing of Longstreet and Hooper and Harris" as a sister node to the VPE head verb "did."

Step 3 would determine that a prepositional phrase exists after both the antecedent's head verb and the VPE and therefore would delete "in his novels and tales" from the antecedent, resulting in the correct antecedent, "produce humorous effects."

This algorithm will correctly handle the 6 cases of subdeletion in the Brown Corpus. However, examples can be constructed for which this algorithm does not account. In the sentence "Julie drove to school on Friday, and Laura did on Saturday," for example, the VPE is "did" and the correct antecedent is "drove to school." In this example, two prepositional phrases--"to school" and "on Friday"--follow the antecedent's head verb "drove." A prepositional phrase, "on Saturday," also exists following the VPE's head verb. Following the above algorithm, both prepositional phrases "to school" and "on Friday" would be deleted, resulting in an incorrect antecedent. The algorithm makes no provisions for cases containing multiple prepositional phrases and noun phrases. Fortunately, such situations seem rare, as none were found in the Brown Corpus.

More significantly, the algorithm also assumes that analogous phrases following the antecedent and VPE always implies subdeletion. That is, it assumes that prepositional phrases or noun phrases following the VPE always implies that like phrases should be deleted from the antecedent. Again, it is possible to imagine a counterexample, for example, "Dad stayed in the Hilton like Mom did in Pittsburgh." Here, the above algorithm would incorrectly remove the prepositional phrase "in the Hilton."

The expectation was that these counter examples would be less frequent than the cases in which the algorithm would correctly remove unwanted text. A manual sampling of VPEs in the Brown Corpus showed this to be true. When the algorithm was implemented, however, the number of correct answers improved to 258, an increase of 1. In addition to solving the 6 cases of subdeletion, the algorithm

introduced 5 errors; each of these new errors involved a noun phrase or prepositional phrase in the VPE that did not require the deletion of a counterpart in the antecedent. For example, one of the newly introduced errors occurred in the fragment "...creaking in the fog as it had for thirty years." The prepositional phrase "for thirty years" in the VPE caused the removal of the phrase "in the fog" from the antecedent, even though the phrases are not parallel in meaning.

These results imply that the structure of a sentence alone is insufficient to detect subdeletion. It is possible, however, that a larger sample of relevant examples would suggest the best choice (to delete or not to delete) in the absence of additional information. Towards these ends, other corpora in the Penn Treebank will be examined with VPEAL. Also, newer versions of the Treebank include semantic tags to adjunct phrases which will aid in preventing the misidentification of subdeletion described above.

## 5. Conclusion

Improving the results of the VPEAL program is an iterative process. We have categorized the errors occurring in VPEAL. An algorithm for solving the error category of subdeletion was described and examined. Potential problem situations for the algorithm were also presented. Empirical evaluation of the algorithm indicates that a purely syntactic approach to detecting subdeletion is probably insufficient. Additional approaches to the problem of subdeletion were suggested. Other cases of errors will be likewise evaluated.

## Bibliography

- Hardt, Daniel. 1995. An empirical approach to VP ellipsis.
- Hobbs, Jerry. 1978. Resolving pronoun references. *Lingua*, 44:311-388.
- Lappin, Shalom and Michael McCord. 1990. Anaphora Resolution in Slot Grammar. *Computational Linguistics*, 16(4).
- Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2).
- Walker, Marilyn. 1989. Evaluating discourse processing algorithms. In *Proceedings, 27th Annual Meeting of the ACL, Vancouver, Canada*.