# An Empirical Approach to VP Ellipsis

Daniel Hardt*
Villanova University

*This paper reports on an empirically based system that automatically resolves VP ellipsis in the 644 examples identified in the parsed Penn Treebank. The results reported here represent the first systematic corpus-based study of VP ellipsis resolution, and the performance of the system is comparable to the best existing systems for pronoun resolution. The methodology and utilities described can be applied to other discourse-processing problems, such as other forms of ellipsis and anaphora resolution.*

*The system determines potential antecedents for ellipsis by applying syntactic constraints, and these antecedents are ranked by combining structural and discourse preference factors such as recency, clausal relations, and parallelism. The system is evaluated by comparing its output to the choices of human coders. The system achieves a success rate of 94.8%, where success is defined as sharing of a head between the system choice and the coder choice, while a baseline recency-based scheme achieves a success rate of 75.0% by this measure. Other criteria for success are also examined. When success is defined as an exact, word-for-word match with the coder choice, the system performs with 76.0% accuracy, and the baseline approach achieves only 14.6% accuracy. Analysis of the individual components of the system shows that each of the structural and discourse constraints used are strong predictors of the antecedent of VP ellipsis.*

## 1. Introduction

Ellipsis is a pervasive phenomenon in natural language, and it has been a major topic of study in theoretical linguistics and computational linguistics in the past several decades (Ross 1967; Sag 1976; Williams 1997; Hankamer and Sag 1976; Webber 1978; Lappin 1984; Sag and Hankamer 1984; Chao 1987; Ristad 1990; Harper 1990; Kitagawa 1991; Dalrymple, Shieber, and Pereira 1991; Lappin 1992; Hardt 1993; Kehler 1993; Fiengo and May 1994). While previous work provides important insight into the abstract syntactic and semantic representations that underlie ellipsis phenomena, there has been little empirically oriented work on ellipsis. The availability of parsed corpora such as the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993) makes it possible to empirically investigate elliptical phenomena in a way not possible before.

This paper reports on an empirically based system that automatically resolves VP ellipsis in the 644 examples identified in the parsed Penn Treebank. This work builds on structural constraints and discourse heuristics first proposed in Hardt (1992) and further developed in Hardt (1995). The results reported here represent the first systematic corpus-based study of VP ellipsis resolution, and the performance of the system is comparable to the best existing systems for pronoun resolution.

The VP elipsis resolution system (VPE-RES) operates on Penn Treebank parse trees to determine the antecedent for VPE occurrences. The system, implemented in Common LISP, uses a Syntactic Filter to eliminate candidate antecedents in impossible

* Department of Computing Science, Villanova University, Villanova, PA 19085

syntactic configurations, and then ranks remaining candidates using Preference Factors involving recency, parallelism, clausal relations, and quotation structure.

All the examples of VPE in the Treebank were coded for the correct antecedent by two coders. Also, a baseline scheme is implemented, which always selects the most recent full VP. Both the system and the baseline are evaluated by comparison with the coder output, with respect to three different definitions of success:

1. **Head Overlap**: either the head verb of the system choice is contained in the coder choice, or the head verb of the coder choice is contained in the system choice
2. **Head Match**: the system choice and coder choice have the same head verb.
3. **Exact Match**: the system choice and coder choice match word-for-word.

Using the Head Overlap measure, the system achieves a success rate of 94.8% on a blind test of 96 Wall Street Journal examples, while the baseline recency scheme achieves a success rate of 75.0% by this measure. Using the Exact Match measure, the system performs with 76.0% accuracy, and the baseline approach achieves 14.6% accuracy.

In what follows, we first present background on the data set and the coding of that data. Next, we describe the VPE-RES system, examining the Syntactic Filter, the Preference Factors, and the Post-Filter. There follows an empirical analysis of the system, in which we compare the system output to coder choice, based on our three success criteria. Also, the subparts of the system are analyzed individually, in three different ways. Finally, we briefly discuss related work.

## 2. Background

In this section, we describe the data set we collected from the Penn Treebank, and the coding of that data.

### 2.1 The Data: VPE in the Penn Treebank

We have identified 644 examples of VPE from the Brown Corpus and the Wall Street Journal Corpus of the Penn Treebank. Since the parsing schemes used in the Penn Treebank do not explicitly label VPE occurrences, it is difficult to ensure that all occurrences of VPE in the Treebank are found. However, this data set is the first large set of ellipsis examples we are aware of, and it provides a solid empirical foundation not only for the current study, but for future research on ellipsis.

We used several techniques to identify ellipsis occurrences in the Treebank, all involving the tree pattern-matching utility *tgrep*.[1] In the Wall Street Journal Corpus, the **-NONE-** category is used to represent a variety of empty expressions, including VPE. We searched for the following pattern:

```
(VP (-NONE- *?*))
```

which resulted in a set of 260 examples from the Wall Street Journal corpus.

1 The utility *tgrep* is written by Rich Pito of the University of Pennsylvania, and is distributed together with the Penn Treebank CD-ROM.

**Table 1**
Identification of VPE occurrences.

| Actual Number | Number Found | False Hits | Recall | Precision |
|---|---|---|---|---|
| 48 | 21 | 19 | 21/48 (44%) | 21/40 (53%) |

**Table 2**
Success rates on missed examples.

| Success Definition | *Missed Examples (27)* Number Correct | *Complete Treebank (644)* Number Correct |
|---|---|---|
| Head Overlap | 25(93%) | 594(92.2%) |
| Head Match | 23(85%) | 537(83.4%) |
| Exact Match | 19(70%) | 489(75.9%) |

**Table 3**
Coder agreement.

| Success Definition | Coder Agreement |
|---|---|
| Head Overlap | 99% |
| Head Match | 97% |
| Exact Match | 93% |

In the Brown Corpus, VPE occurrences are not labeled in this way. We searched for occurrences of a sentence (S) with an auxiliary (AUX) but no VP.

To evaluate our identification criteria, we performed a manual search for VPE occurrences in a sample of files constituting about 3.2% of the Treebank.[2] In this sample, we found that the Recall was 44%, and the Precision was 53%, as depicted in Table 1.

In this sample, there were 27 valid VPE occurrences that were missed. We tested VPE-RES on these examples, and found that its performance was comparable to its performance on the examples that were automatically identified. The results are given in Table 2, which also includes results on the complete corpus, for ease of comparison.[3]

### 2.2 Coding the Data

All the examples of VPE were coded for the correct antecedent by two human coders. We performed some comparisons of coder responses with one another, based on our three success criteria. On a sample of 162 examples, the results were as shown in Table 3.

In this sample, there was only one example in which coder agreement failed according to the Head Overlap criterion:

(1) Gold still acts as a haven when uncertainty prevails in the financial markets as it did yesterday.

2 This percentage was computed using the parsed version of both the Brown Corpus and Wall Street Journal Corpus. The sample contained 155,000 words, out of a total of approximately 4,799,845 in the Treebank.

3 The numbers for the missed examples reflect a post hoc analysis of the program output, rather than comparison with coder files.

Here are the two choices:

Coder 1: prevails in the financial markets

Coder 2: acts as a haven

The following is an example where the coders disagreed according to Head Match, although they agreed according to Head Overlap:

(2) By contrast, in 19th-century Russia, an authoritarian government owned the bank and had the power to revoke payment whenever it chose, much as it would in today's Soviet Union.

Coder 1: revoke payment whenever it chose

Coder 2: owned the bank and had the power to revoke payment whenever it chose

In the following example, the coders disagreed according to Exact Match, although they agreed according to the other two success criteria:

(3) When bank financing for the buy-out collapsed last week, so did UAL's stock.

Coder 1: collapsed

Coder 2: collapsed last week

## 3. VPE-RES System

The VPE-RES system has the following subparts:

1. Syntactic Filter
2. Preference Factors
3. Post-Filter

The candidates for VPE antecedents are all full VPs appearing within a three sentence window—the current sentence and the two preceding sentences.[4] The Syntactic Filter eliminates all VPs that contain the VPE in an improper fashion. A preference ordering is imposed upon the remaining candidate antecedents, based on recency, clausal relations, parallelism, and quotation structure. After the candidates have been weighted according to these Preference Factors, the highest-rated candidate is selected, and its form is modified by a Post-Filter.

4 The limitation to three sentences is arbitrary. However, no examples were found in the Treebank in which the antecedent was more distant.

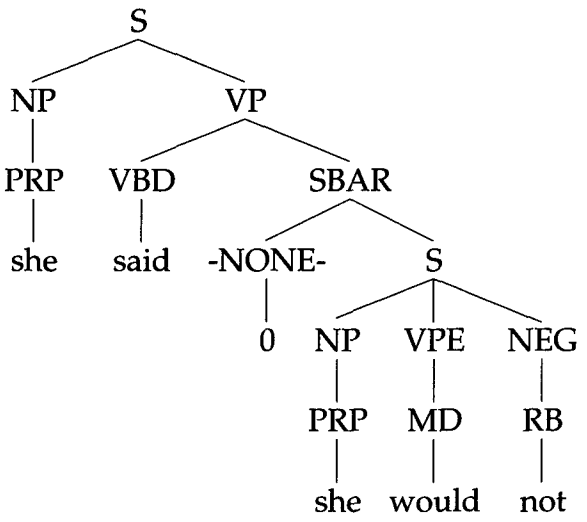


**Figure 1**
Parse tree for *She said she would not.*

### 3.1 Syntactic Filter

The Syntactic Filter rules out antecedents that improperly contain the VPE occurrence.[5] While the precise definition of improper containment is an active area of theoretical research,[6] we rule out antecedents that contain the VPE in a *sentential complement*. An example of this is given in Figure 1, the parse tree for the sentence in (4).[7]

(4) She said she would not.

Here, the VPE occurrence *would* cannot select as its antecedent the containing VP headed by *said*. This is ruled out by the Syntactic Filter, because the VPE is contained in SBar, a sentential complement to *said*.

Pronoun resolution systems often incorporate a syntactic filter—a mechanism to remove certain antecedents based on syntactic structure. The basic syntactic constraint for pronouns is that they cannot take a "local" antecedent, as described, for example, in Principle B of the binding theory (Chomsky, 1981).[8] The Syntactic Filter for VPE also rules out "local" antecedents in a sense: it rules out antecedents in certain containment configurations.

The implementation of the Syntactic Filter is complicated by two factors: first, there are certain cases in which a containing antecedent *is* possible, where the VPE is

5 This constraint is discussed in Hardt (1992) as a way of ruling out antecedents for VPE.

6 See, for example Sag (1976) and May (1985) for discussion, and for example Lappin and McCord (1990) and Jacobson (1992) for alternative views.

7 Parse trees display the exact category labels and structure represented in the Penn Treebank parses. We have added a label, *VPE*, for VPE occurrences. See Appendix A for a list of Penn Treebank tags; for more information, see Marcus, Santorini, and Marcinkiewicz (1993).

8 While the precise formulation of Principle B remains controversial, it is generally agreed to rule out, for example, the binding of a pronoun in object position by an NP in subject position. Such constraints on pronoun resolution have been incorporated into several computational approaches to pronoun resolution, such as Brennan, Friedman, and Pollard (1987), Lappin and McCord (1990), and Lappin and Leass (1994).

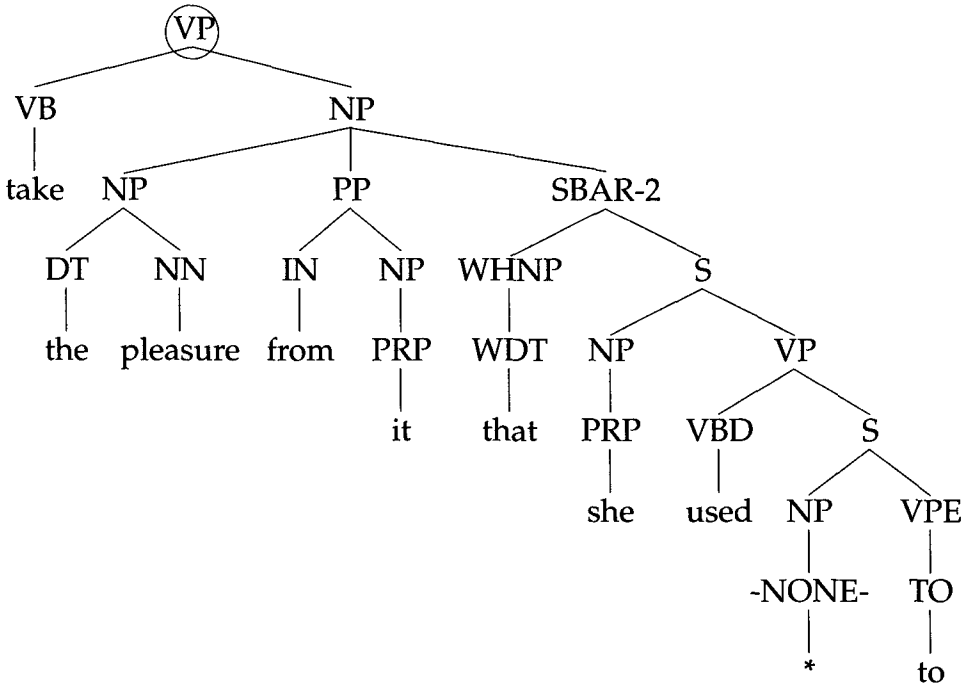


**Figure 2**
Parse tree for *She was getting too old to take the pleasure from it that she used to.*

contained in an NP argument of the containing VP, as in Figure 2, the parse tree for the following example:

(5) She was getting too old to take the pleasure from it that she used to.

Here, the (circled) VP headed by *take* is the antecedent for the VPE, despite the containment relation.

The second complication results from a basic limitation in Treebank parses; there is no distinction between arguments and adjuncts. A VP must be ruled out if the VPE is within a nonquantificational argument; when a VPE occurs in an adjunct position, the "containing" VP is a permissible antecedent. The following sentence, whose parse tree is in Figure 3, is an example of this:

(6) get to the corner of Adams and Clark just as fast as you can

In this case, the (circled) VP headed by *get* is the antecedent for the VPE, despite the appearance of containment. Since the VPE is contained in an adjunct (an adverbial phrase), there is in fact a nonmaximal VP headed by *get* that does not contain the VPE: this is the VP *get to the corner of Adams and Clark*. However, because of the approach taken in annotating the Penn Treebank, this nonmaximal VP is not displayed as a VP.

To capture the above data, the Syntactic Filter rules out VPs that contain the VPE in a sentential complement; any other antecedent-containment relation is permitted.

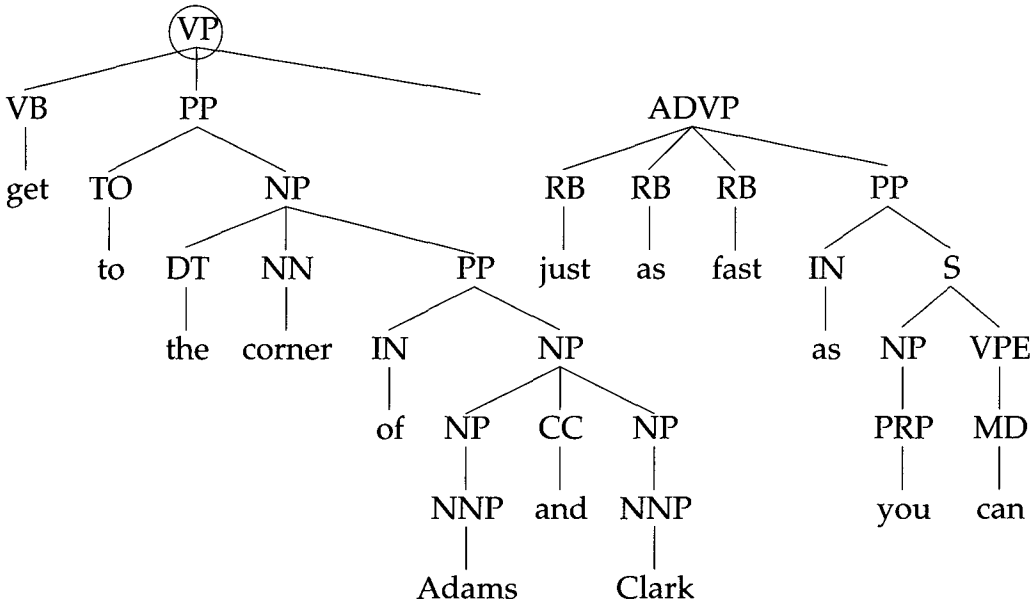


**Figure 3**
Parse tree for *get to the corner of Adams and Clark just as fast as you can.*

This correctly rules out the containing antecedent in (4), and permits it in (5) and (6).[9]

### 3.2 Preference Factors

Remaining candidates are ordered according to the following four Preference Factors:

1. Recency
2. Clausal Relations
3. Parallelism
4. Quotation

9 An anonymous CL reviewer suggests that the filter may be overly restrictive, because of examples like the following:

A: It's an important issue, and I'm very concerned about it.

B: Well, frankly, *I* don't care that *you* do.

(Italicized expressions receive pitch accents.) Here, the antecedent for the VPE is *care*; this would not be permitted by the filter. The reviewer suggests that examples like this should not be categorically excluded, although they are perhaps less than fully acceptable. If this is true, it raises interesting theoretical issues about the acceptability of antecedent-containment configurations. However, the reviewer notes that "such examples are no doubt rare and perhaps the proposed containment filter does enough work in correctly excluding ill-formed instances of ellipsis to justify the categorical exclusion of these cases." Based on our empirical research up to this point, we concur with this. No examples of this sort have been observed among the 644 VPE examples in the Penn Treebank, and the Syntactic Filter as currently formulated contributes significantly to the overall performance of the system (see Section 4 for figures on this).

Each candidate is initialized with a **weight** of 1. This weight is modified by any applicable Preference Factors.

**3.2.1 Recency.** The simplest and most important factor is recency: if no other Preference Factors obtain, the most recent (syntactically possible) antecedent is always chosen. The weights are modified as follows: the first VP weight is set to be the **recency factor**, 1.15. Moving rightward, toward the VPE, the weight of each subsequent VP is multiplied by the **recency factor**. Thus, if there are three VPs preceding the VPE, we have (1.15 1.32 1.52). If a VP contains another VP, the two VPs are set at the same level. Finally, VPs *following* the VPE are penalized in a symmetrical fashion.[10]

**3.2.2 Clausal Relations.** There is a strong preference for a VP antecedent that is in a clausal relation to the VPE.[11] Consider the following example:

(7) tells you what the characters are thinking and feeling [$_{\text{ADVP}}$ far more precisely than intertitles, or even words, [$_{\text{VPE}}$ would]].

The VP headed by *tells* is modified by the adverbial phrase (labeled ADVP) containing the VPE. This VP is the correct antecedent. A VP in such a relation is given a very high weight, by the Preference Factor Clause-Rel, which in practice makes it an obligatory antecedent. If Clause-Rel is deactivated, the system incorrectly selects *feeling* as the antecedent, because it is the most recent VP.

The modification relation can also be a comparative relation, as illustrated by the following example, whose parse tree is given in Figure 4:

(8) All felt freer to discuss things than students had previously.

Here, the correct antecedent is the (circled) VP headed by *felt*. This VP is modified by the comparative clause containing the VPE, and thus is correctly selected by the system. With Clause-Rel deactivated, the system incorrectly selects the more recent VP *discuss things*.

Note that such VPs are parsed as containing the VPE, but they are not removed by the Syntactic Filter. Thus, the effect of this constraint is best observed in conjunction with the Syntactic Filter. In the testing of the system, we examined each system component separately, as described below. However, we also examined Clause-Rel in combination with the Syntactic Filter, because of their close connection. We did this by defining a Composite system component, consisting of Syntactic Filter, Clause-Rel, and Post-Filter.

**3.2.3 Parallelism.** There is a preference for similar parallel elements, that is, the elements surrounding the ellipsis site, and the elements that correspond to them surrounding the antecedent. Notions of parallelism figure prominently in many theoretical studies of ellipsis.[12] However, the proposal that similarity of parallel elements can be

10 This reflects the fact that VPE, like pronominal anaphora, permits the antecedent to follow, rather than precede, the VPE occurrence.

11 This constraint is discussed in Hardt (1992).

12 The term parallel elements is from Dalrymple, Shieber, and Pereira (1991), where parallelism is emphasized in the interpretation of ellipsis. Parallelism is also important in many other treatments of ellipsis, such as Prüst, Scha, and van den Berg (1991), Asher (1993), and Fiengo and May (1994).

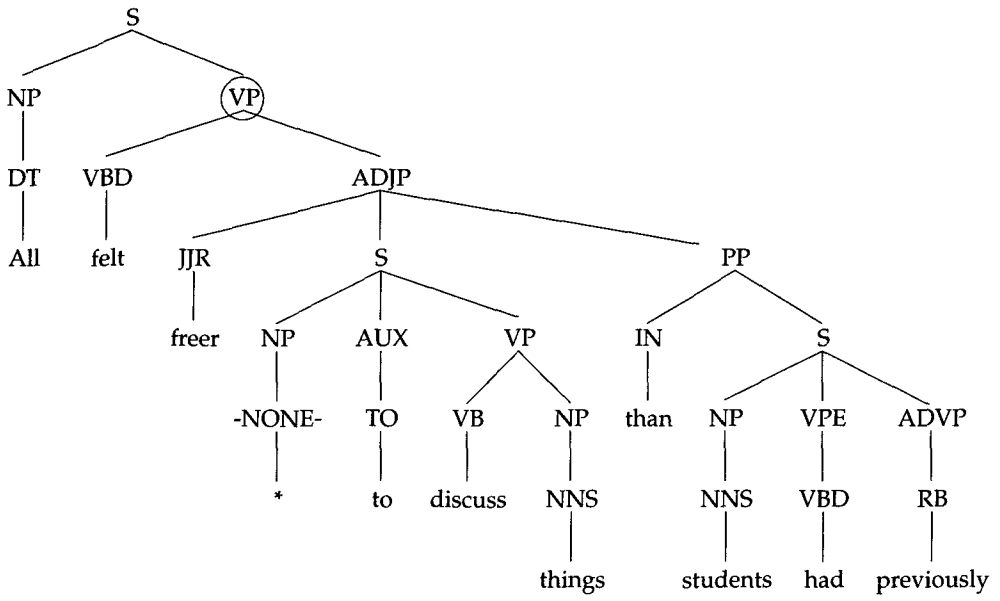


**Figure 4**
Parse tree for *All felt freer to discuss things than students had previously.*

used to guide ellipsis resolution is, to our knowledge, a new one.[13] Our current results involving parallelism provide support for this claim.[14] We are continuing to experiment with more sophisticated ways of measuring the similarity of parallel elements.

In the case of VPE, the subject and auxiliary are parallel elements. Currently, the system only examines the form of the auxiliary. In Hardt (1992) a preference for VPE with coreferential subjects is suggested. This information is not available in the Penn Treebank, and we do not use any forms of subject matching in the current version of the system.

*Aux-Match (Form of Auxiliary).* There is a preference for a similar **base form** of auxiliary in antecedent and VPE. The categories for auxiliary forms we use are: *do, be, have, can, would, should, to.* We prefer an antecedent that shares the same category of auxiliary form as the VPE. The weights of all potential antecedents that do not match the VPE auxiliary category are multiplied by our Standard Penalty Value, which is .667.

This preference is illustrated by the following example:

(9) Someone with a master's degree in classical arts who works in a deli would [$_{VP}$ be ideal], litigation sciences [$_{VP}$ advises]. So [$_{VPE}$ would] someone recently divorced or widowed.

13 The importance of similar parallel elements in discourse relations is emphasized in Hobbs (1979), and it is applied to VPE resolution in Hobbs and Kehler (1997), in a rather different context than that of this paper.

14 As discussed in Section 4, the Parallelism Preference Factor makes an important contribution to the system performance.

Here, the correct antecedent is *be ideal*. It is selected because it has a *would* auxiliary, which is the same category as the VPE. Without this constraint, the system incorrectly selects the VP *advises*.

Another example is the following:

(10) In the past, customers had to [$_{VP}$ go to IBM when they [$_{VP}$ outgrew the Vax]]. Now they don't have [$_{VPE}$ to].

Here, the correct antecedent is the matrix VP headed by *go*. It has a *to* auxiliary, as does the VPE. Without this constraint, the VP *outgrew the VAX* is incorrectly selected by the system.

*Parallel-Match* (Be-do *conflict*). There is an additional penalty for a VP antecedent with a *be*-form auxiliary, if the VPE is a *do*-form.[15] This is implemented by multiplying the VP by our Standard Penalty Value of .667. Consider the following example:

(11) You [$_{VP}$ know what the law of averages [$_{VP}$ is]], don't you VPE?

Here, neither potential antecedent matches the auxiliary category of the VPE, and therefore both are penalized by the general auxiliary match constraint. However, the nearer antecedent, *is*, is a *be*-form, and is thus subject to an additional penalty. This allows the matrix antecedent, *know what the law of averages is*, to be correctly selected.

**3.2.4 Quotation.** If the VPE occurs within quoted material, there is a preference for an antecedent that also occurs within quoted material.[16] This is illustrated by the following example:

(12) "We [$_{VP}$ have good times]." This happy bulletin [$_{VP}$ convulsed Mr. Gorboduc]. "You [$_{VPE}$ do] ? ", he asked between wheezes of laughter.

Here, the correct antecedent is *have good times*. The VP *convulsed Mr. Gorboduc* is penalized by the Standard Penalty Value, because it is not within quotations, while the VPE is within quotations. Without the application of the quote preference, the system incorrectly selects *convulsed Mr. Gorboduc*.

### 3.3 Post-Filter

Once the highest-rated antecedent has been identified, it may be necessary to modify it by removing an argument or adjunct that is incorrectly included. If the selected VP contains the VPE in an argument or adjunct, that argument or adjunct must be eliminated. For example,

(13) Different as our minds are, yours has [$_{VP}$ nourished mine [$_{PP}$ as no other social influence [$_{VPE}$ has ]]].

The antecedent VP selected is *nourished mine as no other social influence ever has*. The PP containing the VPE must be eliminated, leaving the correct antecedent *nourished mine*. This Preference Factor is extremely important in achieving success by the

15 This constraint is suggested in Hardt (1992).
16 A preference of this sort is discussed in Malt (1984).

Exact-Match criterion, and it results in a great deal of improvement over the baseline approach (see results in Section 4).

## 4. Empirical Evaluation

### 4.1 Success Criteria

To test the performance of the system, we first obtained a coded file, which indicates a human coder's preferred antecedent for each example. Then we compared the output of the system with the coder's selections.

As mentioned in Section 1, we define three criteria for success:

1. **Head Overlap**: either the head verb of the system choice is contained in the coder choice, or the head verb of the coder choice is contained in the system choice.
2. **Head Match**: the system choice and coder choice have the same head verb.
3. **Exact Match**: the system choice and coder choice match word-for-word.

To illustrate these criteria, we give three examples, one for each success criterion. Note that the success criteria are increasingly strict—if an example satisfies Exact Match, it will also satisfy the other two criteria, and if an example satisfies Head Match, it will also satisfy Head Overlap.

**Example: Head Overlap**

(14) In July, Par and a 60% owned unit agreed to plead guilty in that inquiry, as did another former Par official.

System output: plead guilty in that inquiry

Coder selection: agreed to plead guilty in that inquiry

According to Head Overlap, the system choice is correct, since its head verb, *plead*, is contained in the coder selection. This would not be considered correct according to Head Match, since the head of the coder selection is *agreed*.

**Example: Head Match**

(15) The question is, if group conflicts still exist, as undeniably they do,

System output: exist

Coder selection: still exist

Here, both the system output and the coder selection have the head verb *exist*, but there is not an exact, word-for-word match.

**Example: Exact Match**

(16) It is difficult if not impossible for anyone who has not pored over the thousands of pages of court pleadings and transcripts to have a worthwhile opinion on the underlying merits of the controversy. Certainly I do not.

**Table 4**
VPE-RES system.

| Success Definition | *Total (644)* Number Correct | *WSJ (260)* Number Correct | *Brown (384)* Number Correct | *Blind Test (96)* Number Correct |
|---|---|---|---|---|
| Head Overlap | 594(92.2%) | 248(95.4%) | 346(90.1%) | 91(94.8%) |
| Head Match | 537(83.4%) | 224(86.2%) | 313(81.5%) | 81(84.4%) |
| Exact Match | 489(75.9%) | 212(81.5%) | 277(72.1%) | 73(76.0%) |

**Table 5**
Baseline (Recency-Only).

| Success Definition | *Total (644)* Number Correct | *WSJ (260)* Number Correct | *Brown (384)* Number Correct | *Blind Test (96)* Number Correct |
|---|---|---|---|---|
| Head Overlap | 495(76.9%) | 196(75.4%) | 299(77.9%) | 72(75.0%) |
| Head Match | 420(65.2%) | 166(63.8%) | 254(66.1%) | 59(61.5%) |
| Exact Match | 188(29.2%) | 52(20.0%) | 136(35.4%) | 14(14.6%) |

System output: have a worthwhile opinion on the underlying merits of the controversy

Coder selection: have a worthwhile opinion on the underlying merits of the controversy

### 4.2 Test Results

After identifying 644 examples of VPE in the Treebank, we reserved 96 randomly selected examples from the Wall Street Journal corpus for a blind test. In Table 4, we give results for the blind test and for the entire Penn Treebank, and we report separate figures on the Brown Corpus and Wall Street Journal Corpus.[17] As a baseline, we also report results (Table 5) on a simple recency-based approach: the most recent VP is always chosen. No Preference Factors or filters are applied.

The difference between the VPE-RES performance and the baseline is statistically significant by all three criteria, based on a $\chi^2$ analysis, $p < .001$.

### 4.3 Evaluating System Subparts

In Tables 6, 7, and 8, we present results on each major subpart of the program. For this evaluation, we used the Exact Match criterion. We evaluated subparts in three ways: first, we began with the baseline (recency) approach, and activated a single additional component, to see how the system performance changed based on that component. Second, we began with the complete system, and deactivated a single component. Finally, we evaluated system components in an incremental fashion, beginning with Post-Filter, then activating Syntactic Filter with Post-Filter still activated, etc. The Composite Factor is a combination of Post-Filter, Syntactic Filter, and Clause-Rel.

17 Since the blind test examples are all taken from the Wall Street Journal corpus, it is most appropriate to compare the blind test results directly to the results on the Wall Street Journal Corpus. Not surprisingly, the blind test results are slightly lower than the results on the complete Wall Street Journal Corpus, since this contains the examples that functioned as training data.

**Table 6**
Recency-only with single factor activated.

| System Subpart | Total (644) Number Correct | WSJ (260) Number Correct | Brown (384) Number Correct | Blind Test (96) Number Correct |
|---|---|---|---|---|
| Recency-Only | 188(29.2%) | 52(20.0%) | 136(35.4%) | 14(14.6%) |
| Post-Filter | 383(59.5%) | 155(59.6%) | 228(59.4%) | 51(53.1%) |
| Syntactic Filter | 232(36.0%) | 73(28.1%) | 159(41.4%) | 21(21.9%) |
| Clause-Rel | 181(28.1%) | 49(18.8%) | 132(34.4%) | 14(14.6%) |
| Quotes | 193(30.0%) | 53(20.4%) | 140(36.5%) | 14(14.6%) |
| Aux-Match | 221(34.3%) | 64(24.6%) | 157(40.9%) | 16(16.7%) |
| Parallel-Match | 201(31.2%) | 56(21.5%) | 145(37.8%) | 14(14.6%) |
| Composite | 461(71.6%) | 200(76.9%) | 261(68.0%) | 71(74.0%) |

**Table 7**
Complete system with single factor de-activated.

| System Subpart | Total (644) Number Correct | WSJ (260) Number Correct | Brown (384) Number Correct | Blind Test (96) Number Correct |
|---|---|---|---|---|
| Full VPE-RES System | 489(75.9%) | 212(81.5%) | 277(72.1%) | 73(76.0%) |
| Post-Filter | 258(40.1%) | 82(31.5%) | 176(45.8%) | 23(24.0%) |
| Syntactic Filter | 431(66.9%) | 185(71.2%) | 246(64.1%) | 61(63.5%) |
| Clause-Rel | 469(72.8%) | 195(75.0%) | 274(71.4%) | 65(67.7%) |
| Quotes | 488(75.8%) | 212(81.5%) | 276(71.9%) | 73(76.0%) |
| Aux-Match | 479(74.4%) | 205(78.8%) | 274(71.4%) | 71(74.0%) |
| Parallel-Match | 479(74.4%) | 209(80.4%) | 270(70.3%) | 73(76.0%) |
| Composite | 232(36.0%) | 67(25.8%) | 165(43.0%) | 16(16.7%) |

**Table 8**
Factors activated incrementally.

| System Subpart | Total (644) Number Correct | WSJ (260) Number Correct | Brown (384) Number Correct | Blind Test (96) Number Correct |
|---|---|---|---|---|
| Recency-Only | 188(29.2%) | 52(20.0%) | 136(35.4%) | 14(14.6%) |
| Post-Filter | 383(59.5%) | 155(59.6%) | 228(59.4%) | 51(53.1%) |
| Syntactic Filter | 445(69.1%) | 187(71.9%) | 258(67.2%) | 63(65.6%) |
| Clause-Rel | 461(71.6%) | 200(76.9%) | 261(68.0%) | 71(74.0%) |
| Quotes | 467(72.5%) | 201(77.3%) | 266(69.3%) | 71(74.0%) |
| Aux-Match | 479(74.4%) | 209(80.4%) | 270(70.3%) | 73(76.0%) |
| Parallel-Match | 489(75.9%) | 212(81.5%) | 277(72.1%) | 73(76.0%) |
| Full VPE-RES System | 489(75.9%) | 212(81.5%) | 277(72.1%) | 73(76.0%) |

### 4.4 System Components

The most important system component is the Composite Factor, which is a combination of the Syntactic Filter, the Post-Filter, and Clause-Rel. The contribution of Clause-Rel is not evident individually; if it is the only factor activated together with Recency-Only, performance in the complete corpus actually declines from 29.2% to 28.1%. However, this is because Clause-Rel requires the Syntactic Filter to make a contribution. This can be observed from the fact that Composite performs better than its individual components. Also, when Clause-Rel is the *deactivated* factor, performance declines from 75.9% to 72.8%. The Parallelism Preference Factors, Aux-Match and Parallel-Match,

also make an important contribution: when they are activated in the incremental analysis, there are 22 additional correct selections in the complete corpus, an improvement of 3.4%.

### 4.5 Errors and Evaluation Criteria

Many of the errors occurring under the Exact Match criterion involve alternatives that are virtually identical in meaning, as in the following example:

(17) Stephen Vincent Benet's *John Brown's Body* [$_{VP}$ comes immediately to mind] [$_{PP}$ in this connection], as does John Steinbeck's *The Grapes Of Wrath* and Carl Sandburg's *The People, Yes.*

Here, VPE-RES selected *comes immediately to mind,* since the PP *in this connection* is parsed as a sister to the VP. One coder selected *comes immediately to mind in this connection,* while the other coder made the same selection as VPE-RES. It is difficult to see any difference in meaning between the two choices.

Because of examples like this, we believe Head Overlap or Head Match are preferable criteria for success. Even with the Head Match criterion, there are errors that involve very subtle differences, such as the following example:

(18) We were there at a moment when the situation in Laos threatened to ignite another war among the world 's giants. Even if it did not, how would this little world of gentle people cope with its new reality of grenades and submachine guns?

The coder selected *ignite another war among the world's giants,* while VPE-RES selected *threatened to ignite another war among the world's giants.*

Some errors result from problems with the Syntactic Filter. The following example illustrates a case of antecedent containment that is not recognized by the filter as currently formulated.

(19) All the generals who held important commands in World War 2, did not write books. It only seems as if they did.

The VPE-RES system incorrectly selects *seems* as the antecedent, because it does not recognize that the VP headed by *seems* improperly contains the VPE.

## 5. Related Work

There is no comparable work we are aware of dealing with VPE resolution; to our knowledge, this is the first empirical study of a VPE resolution algorithm. There is, however, a large body of empirically oriented work on pronoun resolution. A prominent recent example is Lappin and Leass (1994), in which a pronoun resolution system is evaluated on 360 examples taken from computer manuals, with a success rate of 86%. This work involves a post hoc evaluation of the system output, and it appears that evaluation is based on Head Match, although this is not discussed explicitly. The VPE-RES system achieves an 84.4% success rate according to Head Match in the Blind Test data from the Wall Street Journal corpus. This compares favorably with Lappin and Leass's result, especially considering that computer manual text is a good deal more restricted than newspaper text. It is also likely that the VPE-RES success rate would be higher using a post hoc evaluation scheme.

Previous work on pronoun resolution (Hobbs 1978, Walker 1989) reports higher success rates. However, these involved hand-tested algorithms on rather small data sets. Lappin and Leass (1994) implemented and tested Hobbs's algorithm, and reported results that were about 4% less than that of Lappin and Leass (1994).

## 6. Conclusions and Future Work

We have described the first empirical study of VP ellipsis resolution, using a data set of 644 examples from the Penn Treebank to develop and test a VPE resolution system. The system performance is comparable to the best existing systems for pronoun resolution.

The Preference Factors in the system were selected and developed in an iterative testing and refinement process. In future work, we will explore the relationship of these factors to more fundamental and general features of discourse interpretation. We suspect that the preferences for clausal relations, parallelism, and quotation structure all involve clues to the underlying discourse structure, reflecting a general preference for configurations where the VPE clause and antecedent clause participate in a discourse relation.[18] A clausal relation is simply an explicit syntactic clue that there is a discourse relation between two clauses, while similarity of parallel elements is another, more indirect clue of a discourse relation, as discussed for example in Hobbs (1979) and Hobbs and Kehler (1997).

We plan to apply the general approach to other discourse-processing problems, such as other forms of ellipsis, and pronoun resolution. We conjecture that suitably generalized versions of the constraints and heuristics in the current system can be applied to a broad range of discourse-processing problems.

## Appendix A: Treebank Tags

In this appendix we include a list of the Penn Treebank part-of-speech tags (Table 9) and syntactic category labels (Table 10), taken from Marcus, Santorini, and Marcinkiewicz (1993).

**Table 9**
The Penn Treebank POS tagset.

| | | | | | |
|---|---|---|---|---|---|
| 1. | CC | Coordinating conjunction | 14. | NNP | Proper noun, singular |
| 2. | CD | Cardinal number | 15. | NNPS | Proper noun, plural |
| 3. | DT | Determiner | 16. | PDT | Predeterminer |
| 4. | EX | Existential *there* | 17. | POS | Possessive ending |
| 5. | FW | Foreign word | 18. | PRP | Personal pronoun |
| 6. | IN | Preposition/subord. conjunction | 19. | PP$ | Possessive pronoun |
| 7. | JJ | Adjective | 20. | RB | Adverb |
| 8. | JJR | Adjective, comparative | 21. | RBR | Adverb, comparative |
| 9. | JJS | Adjective, superlative | 22. | RBS | Adverb, superlative |
| 10. | LS | List item marker | 23. | RP | Particle |
| 11. | MD | Modal | 24. | SYM | Symbol (mathematical or scientific) |
| 12. | NN | Noun, singular or mass | 25. | TO | *to* |
| 13. | NNS | Noun, plural | 26. | UH | Interjection |

18 For example, Asher (1993) claims that VPE requires a discourse relation between VPE and antecedent clauses.

**Table 9**
Continued.

| | | | |
|---|---|---|---|
| 27. | VB | Verb, base form | |
| 28. | VBD | Verb, past tense | |
| 29. | VBG | Verb, gerund/present participle | |
| 30. | VBN | Verb, past participle | |
| 31. | VBP | Verb, non-3rd ps. sing. present | |
| 32. | VBZ | Verb, 3rd ps. sing. present | |
| 33. | WDT | *wh*-determiner | |
| 34. | WP | *wh*-pronoun | |
| 35. | WP$ | Possessive *wh*-pronoun | |
| 36. | WRB | *wh*-adverb | |
| 37. | # | Pound sign | |
| 38. | $ | Dollar sign | |
| 39. | . | Sentence-final punctuation | |
| 40. | , | Comma | |
| 41. | : | Colon, semi-colon | |
| 42. | ( | Left bracket character | |
| 43. | ) | Right bracket character | |
| 44. | " | Straight double quote | |
| 45. | ` | Left open single quote | |
| 46. | " | Left open double quote | |
| 47. | ' | Right close single quote | |
| 48. | " | Right close double quote | |

**Table 10**
The Penn Treebank syntactic tagset.

Tags

| | | |
|---|---|---|
| 1. | ADJP | Adjective phrase |
| 2. | ADVP | Adverb phrase |
| 3. | NP | Noun phrase |
| 4. | PP | Prepositional phrase |
| 5. | S | Simple declarative clause |
| 6. | SBAR | Clause introduced by subordinating conjunction or *0* (see below) |
| 7. | SBARQ | Direct question introduced by *wh*-word or *wh*-phrase |
| 8. | SINV | Declarative sentence with subject-aux inversion |
| 9. | SQ | Subconstituent of SBARQ excluding *wh*-word or *wh*-phrase |
| 10. | VP | Verb phrase |
| 11. | WHADVP | *Wh*-adverb phrase |
| 12. | WHNP | *Wh*-noun phrase |
| 13. | WHPP | *Wh*-prepositional phrase |
| 14. | X | Constituent of unknown or uncertain category |

Null elements

| | | |
|---|---|---|
| 1. | * | "Understood" subject of infinitive or imperative |
| 2. | 0 | Zero variant of *that* in subordinate clauses |
| 3. | T | Trace—marks position where moved *wh*-constituent is interpreted |
| 4. | NIL | Marks position where preposition is interpreted in pied-piping contexts |

**Acknowledgments**
This work was partially supported by a Villanova University Summer Research Grant, and National Science Foundation Career Grant IRI-9502257. Thanks to Gregg Davis, who did a substantial amount of LISP coding, improving the system performance greatly. Thanks also to Gar Donecker for invaluable help in analysis, coding and organization. Michael Feeney implemented comparison and extraction utilities, and also did coding. Rebecca Passoneau first suggested the use of coders, and also contributed some coding. Bonnie Webber, Mark Steedman, Aravind Joshi, and Marilyn Walker provided important advice and feedback, and useful and constructive suggestions were given by three anonymous CL reviewers.

**References**


Asher, Nicholas. 1993. *Reference to Abstract Objects in English*. Dordrecht.

Brennan, Susan E., Marilyn Walker Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting*. Association for Computational Linguistics.

Chao, Wynn. 1987. *On Ellipsis*. Ph.D. thesis, University of Massachusetts-Amherst.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris.

Dalrymple, Mary, Stuart Shieber, and Fernando Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4).

Fiengo, Robert and Robert May. 1994. *Indices and Identity*. MIT Press, Cambridge, MA.

Hankamer, Jorge and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic Inquiry*, 7(3).

Hardt, Daniel. 1992. An Algorithm for VP Ellipsis. In *Proceedings of the 30th Annual Meeting*, Newark, DE. Association for Computational Linguisitics.

Hardt, Daniel. 1993. *Verb Phrase Ellipsis: Form, Meaning, and Processing*. Ph.D. thesis, University of Pennsylvania.

Hardt, Daniel. 1995. An empirical approach to VP ellipsis. In *Proceedings, AAAI Symposium on Empirical Approaches in Discourse and Generation*, Palo Alto, CA.

Harper, Mary Patricia. 1990. *The Representation of Noun Phrases in Logical Form*. Ph.D. thesis, Brown University.

Hobbs, Jerry. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Hobbs, Jerry. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

Hobbs, Jerry and Andrew Kehler. 1997. A Theory of Parallelism and the Case of VP ellipsis. In *Proceedings of the 35th Annual Meeting*, Madrid, Spain. Association for Computational Linguisitics.

Jacobson, Pauline. 1992. Antecedent contained deletion in a variable-free semantics. In *Proceedings of the Second Conference on Semantics and Linguistic Theory*, Columbus, Ohio.

Kehler, Andrew. 1993. The effect of establishing coherence in ellipsis and anaphora resolution. In *Proceedings of the 31st Annual Meeting*, Columbus, OH. Association for Computational Linguistics.

Kitagawa, Yoshihisa. 1991. Copying identity. *Natural Language and Linguistic Theory*, 9(3):497–536.

Lappin, Shalom. 1984. VP anaphora, quantifier scope, and logical form. *Linguistic Analysis*, 13(4):273–315.

Lappin, Shalom. 1992. The syntactic basis of ellipsis resolution. In *Proceedings of the Stuttgart Ellipsis Workshop*, Stuttgart, Germany.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*.

Lappin, Shalom and Michael McCord. 1990. Anaphora resolution in slot grammar. *Computational Linguistics*, 16(4).

Malt, Barbara. 1984. The role of discourse structure in understanding anaphora. *Journal of Memory and Language*, 24:271–289.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2).

May, Robert. 1985. *Logical Form: Its Structure and Derivation*. MIT Press, Cambridge, MA.

Prüst, Hub, Remko Scha, and Martin van den Berg. 1991. A discourse perspective on verb phrase anaphora. *Linguistics and Philosophy*, 17(3):261–327.

Ristad, E. S. 1990. *Computational Structure of Human Language*. Ph.D. thesis, Massachusetts Institute of Technology.

Ross, H. 1967. Constraints on variables in syntax. MIT Department of Linguistics and Philosophy.

Sag, Ivan and Jorge Hankamer. 1984. Towards a theory of anaphoric processing. *Linguistics and Philosophy*, 7:325–345.

Sag, Ivan A. 1976. *Deletion and Logical Form*. Ph.D. thesis, Massachusetts Institute of Technology. (Published 1980 by Garland Publishing, New York).

Walker, Marilyn. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting*, Vancouver, Canada. Association for Computational Linguistics.

Webber, Bonnie Lynn. 1978. *A Formal Approach to Discourse Anaphora*. Ph.D. thesis, Harvard University. (Published 1979 by Garland Publishing, New York).

Williams, Edwin. 1977. Discourse and logical form. *Linguistic Inquiry*, 8(1):101–139.