The Notion of Argument in Prepositional Phrase Attachment

Paola Merlo*
University of Geneva

Eva Esteve Ferrer†
University of Sussex

In this article we refine the formulation of the problem of prepositional phrase (PP) attachment as a four-way disambiguation problem. We argue that, in interpreting PPs, both knowledge about the site of the attachment (the traditional noun–verb attachment distinction) and the nature of the attachment (the distinction of arguments from adjuncts) are needed. We introduce a method to learn arguments and adjuncts based on a definition of arguments as a vector of features. In a series of supervised classification experiments, first we explore the features that enable us to learn the distinction between arguments and adjuncts. We find that both linguistic diagnostics of argumenthood and lexical semantic classes are useful. Second, we investigate the best method to reach the four-way classification of potentially ambiguous prepositional phrases. We find that whereas it is overall better to solve the problem as a single four-way classification task, verb arguments are sometimes more precisely identified if the classification is done as a two-step process, first choosing the attachment site and then labeling it as argument or adjunct.

1. Motivation

Incorrect attachment of prepositional phrases (PPs) often constitutes the largest single source of errors in current parsing systems. Correct attachment of PPs is necessary to construct a parse tree that will support the proper interpretation of constituents in the sentence. Consider the timeworn example

(1) I saw the man with the telescope.

It is important to determine if the PP *with the telescope* is to be attached as a sister to the noun *the man*, restricting its interpretation, or if it is to be attached to the verb, thereby indicating the instrument of the main action described by the sentence. Based on examples of this sort, recent approaches have formalized the problem of disambiguating PP attachments as a binary choice, distinguishing between attachment of a PP to a given verb or to the verb's direct object (Hindle and Rooth 1993; Ratnaparkhi, Reynar, and Roukos 1994; Collins and Brooks 1995; Merlo, Crocker, and Berthouzoz 1997; Stetina and Nagao 1997; Ratnaparkhi 1997; Zhao and Lin 2004).

This is, however, a simplification of the problem, which does not take the nature of the attachment into account. Precisely, it does not distinguish PP arguments from

* Linguistics Department, University of Geneva, 2 rue de Candolle, 1211 Genève 4, Switzerland.
† Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, UK.

Submission received: 28 November 2003; revised submission received: 22 June 2005; accepted for publication: 4 November 2005.

PP adjuncts. Consider the following example, which contains two PPs, both modifying the verb.

(2) Put the block on the table in the morning.

The first PP is a locative PP required by the subcategorization frame of the verb *put*, whereas *in the morning* is an optional descriptor of the time at which the action was performed. Although both are attached to the verb, the two PPs entertain different relationships with the verb—the first is an argument whereas the latter is an adjunct. Analogous examples could be built for attachments to the noun. (See examples 7a, b.)

Thus, PPs cannot only vary depending on the site to which they attach in the structure, such as in example (1), but they can fulfill different functions in the sentence, such as in example (2). In principle, then, a given PP could be four-way ambiguous. In practice, it is difficult and moderately unnatural to construct examples of four-way ambiguous sentences, sentences that only a good amount of linguistic and extralinguistic knowledge can disambiguate among the noun-attached and verb-attached option, with an argument or adjunct interpretation. It is, however, not impossible.

Consider benefactive constructions, such as the sentence below.

(3) Darcy baked a cake for Elizabeth.

In this case the *for* is a benefactive, hence an argument of the verb *bake*. However, the *for*-PP is optional; thus other non-argument PPs can occur in the same position.

(4) Darcy baked a cake for 5 shillings/for an hour.

Whereas in sentence (3) the PP is an argument, in (4) the PP is an adjunct, as indicated by the different status of the corresponding passive sentences and by the ordering of the PPs (arguments prefer to come first), as shown in (5) and (6).

(5a) Elizabeth was baked a cake by Darcy

(5b) *5 shillings/an hour were baked a cake by Darcy

(6a) Darcy baked a cake for Elizabeth for 5 shillings/for an hour

(6b) ??Darcy baked a cake for 5 shillings/for an hour for Elizabeth

This kind of ambiguity also occurs in sentences in which the *for*-PP is modifying the object noun phrase. Depending on the head noun in object position, and under the assumption that a beneficiary is an argument, as we have assumed in the sentences above, the PP will be an argument or an adjunct, as in the following examples, respectively.

(7a) Darcy baked [cakes for children]

(7b) Darcy baked [cakes for 5 shillings]

Modeling both the site and the nature of the attachment of a PP into the tree structure is important. Distinguishing arguments from adjuncts is key to identifying the elements that belong to the semantic kernel of a sentence. Extracting the kernel of a sentence or phrase, in turn, is necessary for automatic acquisition of important lexical knowledge, such as subcategorization frames and argument structures, which is used in several natural language processing (NLP) tasks and applications, such as parsing, machine translation, and information extraction (Srinivas and Joshi 1999; Dorr 1997; Phillips and Riloff 2002). This task is fundamentally syntactic in nature and complements the task of assigning thematic role labels (Gildea and Jurafsky 2002; Nielsen and Pradhan 2004; Xue and Palmer 2004; Swier and Stevenson 2005). See also the common task of CoNNL (2004, 2005) and SENSEVAL-3 (2004). Both a distinction of arguments from adjuncts and an appropriate thematic labeling of the complements of a predicate, verb, or noun are necessary, as confirmed by the annotations adopted by current corpora. Framenet makes a distinction between complements and satellites (Baker, Fillmore, and Lowe 1998). The developers of PropBank integrate the difference between arguments and adjuncts directly into the level of specificity of their annotation. They adopt labels that are common across verbs for adjuncts. They inherit these labels from the Penn Treebank annotation. Arguments are annotated instead with labels specific to each verb (Xue 2004; Palmer, Gildea, and Kingsbury 2005).

From a quantitative point of view, arguments and adjuncts have different statistical properties. For example, Hindle and Rooth (1993) clearly indicate that their lexical association technique performs much better for arguments than for adjuncts, whether the attachment is to the verb or to the noun.

Researchers have abstracted away from this distinction, because identifying arguments and adjuncts is a notoriously difficult task, taxing many native speakers' intuitions. The usual expectation has been that this discrimination is not amenable to a corpus-based treatment. In recent preliminary work, however, we have succeeded in distinguishing arguments from adjuncts using corpus evidence (Merlo and Leybold 2001; Merlo 2003). Our method develops corpus-based statistical correlates for the diagnostics used in linguistics to decide whether a PP is an argument or an adjunct. A numerical vectorial representation of the notion of argumenthood is provided, which supports automatic classification. In the current article, we expand and improve on this work, by developing new measures and refining the previous ones. We also extend that work to attachment to nouns. This extension enables us to explore in what way the distinction between argument and adjunct is best integrated in the traditional attachment disambiguation problem.

We treat PP attachment as a four-way classification of PPs into noun argument PPs, noun adjunct PPs, verb argument PPs, and verb adjunct PPs. We investigate this new approach to PP attachment disambiguation through several sets of experiments, testing different hypotheses on the argument/adjunct distinction of PPs and on its interaction with the disambiguation of the PP attachment site. The two main claims can be formulated as follows.

- Hypothesis 1: The argument/adjunct distinction can be performed based on information collected from a minimally annotated corpus, approximating deeper semantic information statistically.
- Hypothesis 2: The learning features developed for the notion of argument and adjunct can be usefully integrated in a finer-grained formulation of the problem of PP attachment as a four-way classification.

To test these two hypotheses, we illustrate our technique to distinguish arguments from adjuncts (Section 2), and we report results on this binary classification (Sections 3 and 4). The intuition behind the technique is that we do not need to represent the distinction between arguments and adjuncts directly, but that the distinction can be indirectly represented as a numerical vector. The feature values in the vector are corpus-based numerical equivalents of the grammaticality diagnostics used by linguists to decide whether a PP is an argument or an adjunct. For example, one of the values in the vector indicates if the PP is optional, whereas another one indicates if the PP can be iterated. Optionality and iterability are two of the criteria used by linguists to determine whether a PP is an argument or an adjunct. In Section 5, we show how this distinction supports a more refined formulation of the problem of PP attachment. We compare two methods to reach a four-way classification. One method is a two-step process that first classifies PPs as attached to the noun or to the verb, and then refines the classification by assigning argument or adjunct status to the disambiguated PPs. The other method is a one-step process that performs the four-way classification directly. We find that the latter has better overall performance, confirming our expectation (Hypothesis 2). In Section 6 we discuss the implications of the results for a definition of the notion of argument and compare our work to that of the few researchers who have attempted to perform the same distinction.

2. Distinguishing Arguments from Adjuncts

Solving the four-way classification task described in the introduction crucially relies on the ability to distinguish arguments from adjuncts, using corpus counts. The ability to automatically make this distinction is necessary for the correct automatic acquisition of important lexical knowledge, such as subcategorization frames and argument structures, which is used in parsing, generation, machine translation, and information extraction (Srinivas and Joshi 1999; Stede 1998; Dorr 1997; Riloff and Schmelzenbach 1998). Yet, few attempts have been made to make this distinction automatically.

The core difficulty in this enterprise is to define the notion of argument precisely enough that it can be used automatically. There is a consensus in linguistics that arguments and adjuncts are different both with respect to their function in the sentence and in the way they themselves are interpreted (Jackendoff 1977; Marantz 1984; Pollard and Sag 1987; Grimshaw 1990). With respect to their function, an argument fills a role in the relation described by its associated head, whereas an adjunct predicates a separate property of its associate head or phrase. With respect to their interpretation, a complement is an argument if its interpretation depends exclusively on the head with which it is associated, whereas it is an adjunct if its interpretation remains relatively constant when associating with different heads (Grimshaw 1990, page 108). These semantic differences give rise to some observable distributional consequences: for a given interpretation, an adjunct can co-occur with a relatively broad range of heads, whereas arguments are limited to co-occurrence with a (semantically restricted) class of heads (Pollard and Sag 1987, page 136).

Restricting the discussion to PPs, these differences are illustrated in the following examples (PP-argument in bold); see also Schütze (1995, page 100).

(8a) Maria is a student **of physics**.

(8b) Maria is a student from Phoenix.

In example (8a), the head *student* implies that a subject is being studied. The sentence tells us only one property of Maria: that she is a student of physics. In example (8b), the PP instead predicates a different property of the student, namely her geographical origin, which is not implied by the head *student*.

(9a) Kim camps/jogs/meditates on Sunday.

(9b) Kim depended/blamed the arson **on Sandy**.

In example (9a) the PP *on Sunday* can be construed without any reference to the preceding part of the sentence, and it preserves its meaning even when combining with different heads. This is, however, not the case for (9b). Here, the PP can only be properly understood in connection with the rest of the sentence: Sandy is the person on whom someone depends or the person on which the arson is blamed.

These semantic distinctions between arguments and adjuncts surface in observable syntactic differences and can be detected automatically both by using general formal features and by specific lexical semantic features, which group together the arguments of a lexical head. Unfortunately, the linguistic diagnostics that are used to determine whether a PP is an adjunct or an argument are not accurate in all circumstances, they often partition the set of the examples differently, and they give rise to relative, and not absolute, acceptability judgments.

We propose a methodology that retains both the linguistic insight of the grammatical tests and the ability to effectively combine several gradient, partial diagnostics, typical of automatic induction methods. Specifically, we first find countable diagnostics for the argument–adjunct distinction, which we approximate statistically and estimate using corpus counts. We also augment the feature vector with information encoding the semantic classes of the input words. The diagnostics and the semantic classes are then automatically combined in a decision tree induction algorithm. The diagnostics are presented below.

2.1 The Diagnostics

Many diagnostics for argumenthood have been proposed in the literature (Schütze 1995). Some of them require complex syntactic manipulation of the sentence, such as *wh*-extraction, and are therefore too difficult to apply automatically. We choose six formal diagnostics that can be captured by simple corpus counts: head dependence, optionality, iterativity, ordering, copular paraphrase, and deverbal nominalization. These diagnostics tap into the deeper semantic properties that distinguish arguments from adjuncts, without requiring that the distinctions be made explicit.

Head Dependence. Arguments depend on their lexical heads because they form an integral part of the phrase. Adjuncts do not. Consequently, PP-arguments can only appear with the specific verbal head by which they are lexically selected, whereas PP-adjuncts can co-occur with a far greater range of different heads than arguments because they are necessary for the correct interpretation of the semantics of the verb, as illustrated in the following example sentences.

(10a) a man/woman/dog/moppet/scarecrow with gray hair

(10b) a menu/napkin/glass/waitress/matchbook from Rosie's

(11a) a member/*dog/*moppet/*scarecrow of Parliament

(11b) a student/*punk/*watermelon/*Martian/*poodle/*VCR of physics

We expect an argument PP to occur with fewer heads, whereas an adjunct PP will occur with more heads, as it is not required by a specific verb, but it can in principle adjoin to any verb or noun head.

We capture this insight by estimating the dispersion of the distribution of the different heads that co-occur with a given PP in a corpus. We expect adjunct PPs to have higher dispersion than argument PPs. We use entropy as a measure of the dispersion of the distribution, as indicated in equation (1) (*h* indicates the noun or verb head to which the PP is attached, *X* is the random variable whose outcomes are the values of *h*).

$$\mathrm{hdep(PP)} \approx H_{\mathrm{PP}}(X) = -\Sigma_{h \in X} p(h) \log_2 p(h) \tag{1}$$

Optionality. In most cases, PP-arguments are obligatory elements of a given sentence whose absence leads to ungrammaticality, while adjuncts do not contribute to the semantics of any particular verb, hence they are optional, as illustrated in the following examples (PP-argument in bold):

(12a) John put the book **in the room**.

(12b) *John put the book.

(12c) John saw/read the book in the room.

(12d) John saw/read the book.

Since arguments are obligatory complements of a verb, whereas adjuncts are not, we expect knowledge of a given verb to be more informative with respect to the probability of existence of an argument than of an adjunct. Thus we expect that the predictive power of a verb with regard to its complements will be greater for arguments than for adjuncts.[1] The notion of optionality can be captured by the conditional probability of a PP given a particular verbal head, as indicated in equation (2).

$$\mathrm{opt(PP)} \approx P(\mathrm{PP}|v) \tag{2}$$

Iterativity and Ordering. Because they receive a semantic role from the selecting verb, arguments of the same type cannot be iterated because verbs can only assign any given type of role once. Moreover, in English, arguments must be adjacent to the selecting lexical head. Neither of these two restrictions apply to adjuncts, which can be iterated, and follow arguments in a sequence of PPs. Consequently, in a sequence of several PPs

[1] Notice that this diagnostic can only be interpreted as a statistical tendency, and not as a strict test, because not all arguments are obligatory (but all adjuncts are indeed optional). The best known descriptive exception to the criterion of optionality is the class of so-called object-drop verbs (Levin 1993). Here a given verb may tolerate the omission of its argument. In other words, a transitive verb, such as *kiss*, can also act like an intransitive. With respect to optional PPs, it has been argued that instrumentals are arguments (Schütze 1995). While keeping these exceptions in mind, we maintain optionality as a valid diagnostic here.

only the first one can be an argument, whereas the others must be adjuncts, as illustrated in the examples below.

(13a) *Chris rented the gazebo **to yuppies, to libertarians**.

(13b) Kim met Sandy in Baltimore in the hotel lobby in a corner.

These two criteria combined give rise to one diagnostic. The probability of a PP being able to iterate, and consequently being an adjunct, can be approximated as the probability of its occurrence in second position in a sequence of PPs, as indicated in equation (3), where we indicate the position of the PP in a sequence as a subscript.

$$\text{iter}(\text{PP}_1) \approx P(\text{PP}_2) \tag{3}$$

Copular Paraphrase. The diagnostic of copular paraphrase is specific to the distinction of NPs arguments and adjuncts, following Schütze (1995, page 103). It does not apply to VP arguments and adjuncts, as it requires paraphrasing the PP with a relative clause. Arguments cannot be paraphrased by a copular relative clause, as the examples in (15) show, whereas adjuncts can, as is shown in (14):

(14)	a. a man from Paris	a man who was from Paris
	b. the albums on the shelf	the albums that were on the shelf
	c. the people on the payroll	the people who were on the payroll
(15)	a. the destruction of the city	*the destruction that was of the city
	b. the weight of the cow	*the weight that was of the cow
	c. a member of Parliament	*a member who was of Parliament

This is because a PP attached to a noun as an argument does not predicate a secondary property of the noun, but it specifies the same property that is indicated by the head. To be able to use the relative clause construction, there must be two properties that are being predicated of the same entity.

Thus, the probability that a PP is able to be paraphrased, and therefore that it is an adjunct, can be approximated by the probability of its occurrence following a construction headed by a copular verb, *be, become, appear, seem, remain* (Quirk et al. 1985), as indicated in equation (4), where $\prec$ indicates linear precedence.

$$\text{para}(\text{PP}) \approx P(v_{\text{copula}} \prec \text{PP}) \tag{4}$$

Deverbal Nouns. This diagnostic is based on the observation that PPs following a deverbal noun are likely to be arguments, as the noun shares the argument structure of the verb.[2] Proper counting of this feature requires identifying a deverbal noun in the head noun position of a noun phrase. We identify deverbal nouns by inspecting their morphology (Quirk et al. 1985). Specifically, the suffixes that can combine

[2] Doubts have been cast on the validity of this diagnostic (Schütze 1995), based on work in theoretical linguistics (Grimshaw 1990). Argaman and Pearlmutter (2002), however, have shown that the argument structures of verbs and related nouns are highly correlated. Hence, we keep deverbal noun as a valid diagnostic here, although we show later that it is not very effective.

with verb bases to form deverbal nouns are listed and exemplified in Figure 1 on page 348. This diagnostic can be captured by a probability indicator function, which assigns probability 1 of being an argument to PPs following a deverbal noun and 0 otherwise.

$$\text{deverb(PP)} = \begin{cases} 1 & \text{if deverbal n} \prec \text{PP} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In conclusion, the diagnostics of head dependence, optionality, iterativity, ordering, copular paraphrase, and deverbal nominalization are promising indicators of the status of PPs as either arguments or adjuncts. In Section 3 we illustrate how they can be quantified in a faithful way and, thanks to their simplicity, how they can be estimated in a sufficiently large corpus by simple counts.

Another class of features is also very important for the distinction between arguments and adjuncts, the lexical semantic class to which the lexical heads belong, as we illustrate below.

Lexical Semantic Class Features. According to Levin (1993), there is a regular mapping between the syntactic and semantic behavior of a verb. This gives rise to a lexicon where verbs that share similar syntactic and semantic properties are organized into classes. More specifically, it is assumed that similar underlying components of meaning give rise to similar subcategorization frames and projections of arguments at the syntactic level. Since an argument participates in the subcategorization frame of the verb, whereas an adjunct does not, we expect the argument-taking properties of verbs to be also organized around semantic classes. We expect, therefore, that knowledge of the class of the verb will be beneficial to the acquisition of the distinction between arguments and adjuncts for an individual verb. For example, all verbs of giving take a dative indirect object and all benefactive verbs can take a benefactive prepositional phrase complement (see examples 16). An analogous prediction can be made for nouns. Unlike the diagnostics features, these lexical features do not have a quantitative counterpart, but they are represented as discrete nominal values that indicate the lexical semantic class the words belong to.

(16a) Darcy offered a gift to Elizabeth.

(16b) Darcy cooked a roast for Elizabeth.

-ANT	inhabitant, contestant, informant, participant, lubricant.
-EE	appointee, payee, nominee, absentee, refugee.
-ER, OR	singer, writer, driver, employer, accelerator, incubator, supervisor.
-AGE	breakage, coverage, drainage, leverage, shrinkage, wastage.
-AL	refusal, revival, dismissal.
-ION	exploration, starvation, ratification, victimization, foundation.
-SION	invasion, evasion.
-ING	building, opening, filling, earnings, savings, shavings, wedding.
-MENT	arrangement, amazement, puzzlement, embodiment, equipment.

Figure 1
Nominal endings that indicate deverbal derivation.

We use all these diagnostics, which in linguistics are used as tests of the argument status of PPs, as a distributed representation of argumenthood itself. We do not assume that the syntactic representation of arguments is different from the representation of adjuncts; for example, we do not assume they have a different attachment configuration, rather the diagnostics themselves determine a multidimensional space in which PPs are positioned with different degrees of argumenthood. For such an approach to work, we need to be able to transform each diagnostic into a symbolic or numeric feature and combine the features in a precise way. In the two following sections we illustrate how to calculate the values of each diagnostic using corpus-based approximations and how to combine them with widely used automatic acquisition techniques.

3. Methodology

The diagnostics described above can be estimated by simple corpus counts. The accuracy of the data collection is key to the success of the classifier induction based on these counts. We explain the details of our methodology below.

3.1 Materials

We construct two corpora comprising examples of PP sequences. A PP is approximated as the preposition and the PP-internal head noun. For example, *with very many little children* will be represented as the bigram *with children*. One corpus contains data encoding information for attachment of single PPs in the form of four head words (verb, object noun, preposition, and PP internal noun) indicating the two possible attachment sites and the most important words in the PP for each instance of PP attachments found in the corpus. We also create an auxiliary corpus of sequences of two PPs, where each data item consists of verb, direct object, and the two following PPs. This corpus is only used to estimate the feature Iterativity. All the data were extracted from the Penn Treebank using the tgrep tools (Marcus, Santorini, and Marcinkiewicz 1993). Our goal is to create a more comprehensive and possibly more accurate corpus than the corpora used by Merlo and Leybold (2001), Merlo, Crocker, and Berthouzoz (1997), and Collins and Brooks (1995), among others. To improve coverage, we extracted all cases of PPs following transitive and intransitive verbs and following nominal phrases. We include passive sentences and sentences containing a sentential object. To improve accuracy, we insured that we did not extract overlapping data, contrary to practice in previous PP corpora construction, where multiple PP sequences were extracted more than once, each time as part of a different structural configuration. For example, in previous corpora, the sequence *using crocidolite in filters in 1956,* which is a sequence of two PPs, is counted both as an example of a two PPs sequence as well as an example of a single PP sequence, *using crocidolite in filters.* This technique of using subsequences as independent examples is used both in the corpora used in Merlo and Leybold (2001) and (Merlo, Crocker, and Berthouzoz 1997), and to an even larger extent in the corpus used in Collins and Brooks (1995), who would also have in their corpus the artificially constructed sequence *using crocidolites in 1956.* This method increases the number of available examples, and it is therefore hoped that it will be beneficial to the learning accuracy. However, as shown in Merlo, Crocker, and Berthouzoz (1997), it rests on the incorrect assumption that the probability of an attachment is independent of the position of the PP to be disambiguated in a sequence of multiple PPs. Therefore, we have decided not to decompose the examples into smaller sequences. The possible grammatical configurations that we have taken into account to construct the corpus are exemplified in Appendix 1.

Once the quadruples constituting the data are extracted from the text corpus, it is necessary to separate the statistics corpus from the training and test corpus.[3] Before illustrating the adopted solution to this problem, let us define the following terms.

- The statistics corpus C_{St} is the part of the corpus that is used to extract the tuples that are used to calculate the features.
- The training corpus C_{Tr} is the part of the corpus that is used to extract the tuples that are used as training data for the classifier.
- The testing corpus C_{Te} is the part of the corpus that used to extract the tuples that are used as testing data to evaluate the classifier.
- The training data S_{Tr} is the set of tuples in C_{Tr}, augmented with the features calculated using C_{St}.
- The testing data S_{Te} is the set of tuples in C_{Te}, augmented with the features calculated using C_{St}.

Note that for the testing data S_{Te} to be an independent test set, the testing corpus C_{Te} must be disjoint both from C_{Tr}, the training corpus, but also it must be disjoint from C_{St}, the corpus on which the statistics are calculated. One possible solution, for example, is to equate the statistics and the training corpus, $C_{\mathrm{St}} = C_{\mathrm{Tr}}$, and to assign them Sections 1–22, while C_{Te} = Section 23, thus making the testing corpus C_{Te} be disjoint from both the statistics and the training corpus. The problem with this solution is that the training data S_{Tr} is no longer extracted using the same process as the testing data S_{Te}, and is therefore not good data from which to generalize. In particular, all tuples in the training data S_{Tr} necessarily also occur in the statistics corpus C_{St}, and therefore no vectors in the training data S_{Tr} involve data unseen in the statistics corpus. In contrast, the testing data S_{Te} can be expected to include tuples that did not also occur in the statistics corpus, and so the classifier might not generalize to these tuples using the features we calculate.

The solution we adopt is to split Sections 1 to 22 into two disjoint subcorpora. Because the Penn Treebank is not uniformly annotated across sections, we do not assign whole sections to either the statistical or the training corpus, but instead randomly assign individual sentences to either corpus. Since data sparseness is a more important issue when calculating the features than it is for training the decision tree, we assigned a larger proportion of the corpus to the statistical subcorpus. We assigned 25% of Sections 1–22 to C_{Tr} and the rest to C_{St}. Section 24 is used as a development corpus and Section 23 is the testing corpus C_{Te}. In this setting, since the statistics, training, and testing corpora are all mutually disjoint, all issues of dependence are resolved, and the training data are representative of the real data we are interested in, so we can expect our classifier to be able to generalize to new data.

[3] We thank Eric Joanis for his help in correctly sampling the corpus and calculating the features.

3.2 The Counts of the Learning Features

As we said above, accurate estimates of the values of the features are crucial for the automatic learning algorithm to be successful. We illustrate the estimation of the features below. Often several ways of estimating the features have been implemented, mostly to address sparseness of data.

Lexical Word Classes. As indicated in the previous section, the head word of the governor of the PP, the noun or the verb, is very directly related to the status of the PP. For example, all giving verbs take a PP introduced by *to*, which is an argument of the verb. The lexical semantic class of the head words is therefore going to be very relevant to the disambiguation task.

The semantic grouping has been done automatically, using the lexicographic classes in WordNet 1.7 (Miller et al. 1990). Nouns are classified in different classes, among which, for example, are *animal, artifact, attribute, body, cognition, communication, event, feeling, food, location, motive, person, plant, process, quantity, relation, shape,* and *substance.* This classification required selecting the most frequent WordNet sense for those polysemous nouns being classified and generalizing to its class.[4]

For all the features below and where appropriate, we assume the following notation. Let h be the head, that is, the verb in the features related to verb attachments and the noun in the features related to noun attachment. Let p be the preposition, and $n2$ be the object of the preposition. Let h_{cl} and $n2_{cl}$ be the WordNet class of h and $n2$, respectively. Let $C(h, p, n2)$ be the frequency with which $p, n2$ co-occurs as a prepositional phrase with h. Let $C(h)$ be the frequency of h.

Head Dependence. Head dependence of a PP on a head is approximated by estimating the dispersion of the PP over the possible heads. In a previous attempt to capture this notion, we approximated by simply measuring the cardinality of the set of heads that co-occur with a given PP in a corpus, as indicated in equation (6). The expectation was that a low number indicated argument status, whereas a high number indicated adjunct status (Merlo and Leybold 2001).

$$\text{hdep}(h, p, n2) = |\{h_1, h_2, h_3, \ldots, h_n\}_{p,n2}| \tag{6}$$

By measuring the cardinality of the set of heads, we approximate the dispersion of the distribution of heads by its range. This is a very rough approximation, as the range of a distribution does not give any information on the distribution's shape. The range of a distribution might have the same value for a uniform distribution or a very skewed distribution. Intuitively, we would like a measure that tells us that the former corresponds to a verb with a much lower head dependence than the latter. Entropy is

[4] The automatic annotation of nouns and verbs in the corpus has been done by matching them with the WordNet database files. Before doing the annotation, though, some preprocessing of the data was required to maximize the matching between our corpus and WordNet. The changes made were inspired by those described in Stetina and Nagao (1997, page 75). To lemmatize the words we used "morpha," a lemmatizer developed by John A. Carroll and freely available at the address: http://www.informatics.susx.ac.uk./research/nlp/carroll/morph.html. Upon simple observation, it showed a better performance than the frequently used Porter Stemmer for this task.

a more informative measure of the dispersion of a distribution, which depends both on the range and on the shape of a distribution.

The head dependence measure based on entropy, then, is calculated as indicated in equation (7), which calculates the entropy of the probability distribution generated by the random variable X, whose values are all the heads that co-occur with a given PP.

$$\mathrm{hdep}(h,p,n2) = H_{p,n2}(X) \approx -\Sigma_{h \in X} \frac{C(h,p,n2)}{\Sigma_i C(h_i,p,n2)} \log_2 \frac{C(h,p,n2)}{\Sigma_i C(h_i,p,n2)} \tag{7}$$

The counts that are used to estimate this measure will depend on finding exactly PPs with the same PP internal noun, and on attaching to exactly the same lexical head. We can expect these measures to suffer from sparse data. We implement then some variants of this measure, where we cluster PPs according to the semantic content of the PP-internal nouns and we cluster nominal heads according to their class. The semantic grouping has been done automatically, as indicated in the paragraph above, on calculating word classes. Since WordNet has a much finer-grained top-level classification for nouns than for verbs, we found that grouping head nouns into classes yielded useful generalizations, but it did not do so for verbs.

Therefore, we calculate head dependency in three different variants: One measure is based on PP-internal noun tokens, another variant is based on noun classes for the PP-internal noun position, and another variant is based on classes for both the PP-internal and the head noun position, as indicated in equation (8).

$$\mathrm{hdep}(h,p,n2) = \begin{cases} H_{p,n2}(X) \approx -\Sigma_{h \in X} \frac{C(h,p,n2)}{\Sigma_i C(h_i,p,n2)} \log_2 \frac{C(h,p,n2)}{\Sigma_i C(h_i,p,n2)}, & \text{or} \\ H_{p,n2_{\mathrm{cl}}}(X) \approx -\Sigma_{h \in X} \frac{C(h,p,n2_{\mathrm{cl}})}{\Sigma_i C(h_i,p,n2_{\mathrm{cl}})} \log_2 \frac{C(h,p,n2_{\mathrm{cl}})}{\Sigma_i C(h_i,p,n2_{\mathrm{cl}})}, & \text{or} \\ H_{p,n2_{\mathrm{cl}}}(X_{\mathrm{cl}}) \approx -\Sigma_{h_{\mathrm{cl}} \in X} \frac{C(h_{\mathrm{cl}},p,n2_{\mathrm{cl}})}{\Sigma_i C(h_{\mathrm{cl},i},p,n2_{\mathrm{cl}})} \log_2 \frac{C(h_{\mathrm{cl}},p,n2_{\mathrm{cl}})}{\Sigma_i C(h_{\mathrm{cl},i},p,n2_{\mathrm{cl}})}, & \text{if h = noun.} \end{cases} \tag{8}$$

Optionality. As explained above, we expect that the predictive power of a verbal head—recall that optionality does not apply to noun attachments—about its complements will be greater for arguments than for adjuncts. This insight can be captured by the conditional probability of a PP given a particular verb, as indicated in equation (9).

$$\mathrm{opt}(v,p,n2) \approx \frac{C(v,p,n2)}{C(v)} \tag{9}$$

Analogously to the measure of head dependence for noun attachments, optionality is measured in three variants. First, it is calculated as a conditional probability based on simple word counts in the corpus of single PPs, as indicated in equation (9) above. Second, we also implement a variant that relies on verb classes instead of individual verbs to address the problem of sparse data. Finally, we also implement a variant that relies on noun classes for the PP-internal noun and verb classes instead of individual verbs. For both these measures, verbs and nouns were grouped into classes using

WordNet 1.7 with the same method as for head dependence. The three measures of optionality are indicated in equation (10).

$$\operatorname{opt}(v,p,n2) = \begin{cases} P(p,n2|v) \approx \frac{C(v,p,n2)}{C(v)}, & \text{or} \\ P(p,n2|v_{\mathrm{cl}}) \approx \frac{C(v_{\mathrm{cl}},p,n2)}{C(v_{\mathrm{cl}})}, & \text{or} \\ P(p,n2_{\mathrm{cl}}|v_{\mathrm{cl}}) \approx \frac{C(v_{\mathrm{cl}},p,n2_{\mathrm{cl}})}{C(v_{\mathrm{cl}})}. \end{cases} \tag{10}$$

Iterativity and Ordering. Iterativity and ordering are approximated by collecting counts indicating the proportion of cases in which a given PP in first position had been found in second position in a sequence of multiple PPs over the total of occurrences in any position, as indicated in equation (11). The problem of sparse data here is especially serious because of the small frequencies of multiple PPs. We addressed this problem by using a backed-off estimation, where we replace lexical items by their WordNet classes and collect counts on this representation. Specifically, the iterativity measure has been implemented as follows.

Let $C_{2nd}(h,p,n2)$ be the frequency with which $p,n2$ occurs as a second prepositional phrase with h, and $C_{any}(h,p,n2)$ be the frequency with which it occurs with h in any position.[5] Then:

$$\operatorname{iter}(h,p,n2) \approx \begin{cases} \dfrac{C_{2nd}(h,p,n2)}{C_{any}(h,p,n2)}, & \text{if } C_{any}(h,p,n2) \neq 0\text{, or else} \\ \dfrac{C_{2nd}(h,p,n2_{\mathrm{cl}})}{C_{any}(h,p,n2_{\mathrm{cl}})}, & \text{if } C_{any}(h,p,n2_{\mathrm{cl}}) \neq 0\text{, or else} \\ \dfrac{C_{2nd}(h_{\mathrm{cl}},p,n2_{\mathrm{cl}})}{C_{any}(h_{\mathrm{cl}},p,n2_{\mathrm{cl}})}. \end{cases} \tag{11}$$

Copular Paraphrase. Copular paraphrase is captured by calculating the proportion of times a given PP is found in a copular paraphrase. We approximate this diagnostic by making the hypothesis that a PP following a nominal head is an adjunct if it is also found following a copular verb, *be, become, appear, seem, remain* (Quirk et al. 1985). We calculate then the proportion of times a given PP follows a copular verb over the times it appears following any verb. This count is an approximation because even when we find a copular verb, it might not be part of a relative clause. Here again, we back off to the noun classes of the PP-internal noun to address the problem of sparse data.

$$\operatorname{para}(h,p,n2) \approx \begin{cases} \dfrac{C(v_{\mathrm{copula}} \prec (p,n2))}{\Sigma_i C(v_i \prec (p,n2))}, & \text{if } C(v_{\mathrm{copula}}) \neq 0\text{, or else} \\ \dfrac{C(v_{\mathrm{copula}} \prec (p,n2_{\mathrm{cl}}))}{\Sigma_i C(v_i \prec (p,n2_{\mathrm{cl}}))}. \end{cases} \tag{12}$$

[5] Note that we approximate the prepositions occurring in any position by looking only at the first two prepositions attached to the verb phrase.

Deverbal Nominalization. The diagnostic of deverbal nouns is implemented as a binary feature that simply indicates if the PP follows a deverbal noun or not.

$$\text{deverb}(n, p, n2) = \begin{cases} 1 \text{ if deverbal n} \prec \text{(p,n2)} \\ 0 \text{ otherwise} \end{cases} \tag{13}$$

Deverbal nouns are identified by inspecting their morphology (Quirk et al. 1985). As our corpus is lemmatized, we are confident that all the nouns in it are in their base forms. The suffixes that can combine with verb bases to form deverbal nouns are shown in Figure 1.

The counts that are collected in the way described above constitute a quantified vector corresponding to a single PP exemplar. These exemplars are the input to an automatic classifier that distinguishes arguments from adjuncts, as described in Section 4. Before we describe the experiments, however, attention must be paid to the method that will be used to determine the target attribute—argument or adjunct—that will be used to train the learner in the learning phase and to evaluate the accuracy of the learned classifier in the testing phase.

3.3 The Target Attribute

Since we are planning to use a supervised learning method, we need to label each example with a target attribute. Deciding whether an example is an instance of an argument or of an adjunct requires making a distinction that the Penn Treebank annotators did not intend to make. The automatic annotation of this attribute therefore must rely on the existing labels for the PP that have been given by the Penn Treebank annotators, inferring from them information that was not explicitly marked. We discuss here the motivation for our interpretation.

The PTB annotators found that consistent annotation of argument status and semantic role was not possible (Marcus et al. 1994). The solution adopted, then, was to structurally distinguish arguments from adjuncts only when the distinction was straightforward and to label only some clearly distinguishable semantic roles. Doubtful cases were left untagged. In the Penn Treebank structural distinctions concerning arguments and adjuncts have been oversimplified: All constituents attached to VP are structurally treated as arguments, whereas all constituents attached to NP are treated as adjuncts. The only exception are the arguments of some deverbal nouns, which are represented as arguments. Information about the distinction between arguments and adjuncts, then, must be gleaned from the semantic and function tags that have been assigned to the nodes. Constituents are labeled with up to four tags (including numerical indices) that account for the syntactic category of the constituent, its grammatical function, and its semantic role (Bies et al. 1995). Figure 2 illustrates the tags that involve PP constituents.

From the description of this set of tags we can already infer some information about the argument status of the PPs. PPs with a semantic tag (LOC, MNR, PRP, TMP) are adjuncts, whereas labels indicating PP complements of ditransitive verbs (BNF, DTV) or locative verbs like *put* are arguments. There are, though, some cases that remain ambiguous and therefore require a deeper study. These are untagged PPs and PPs tagged -CLR. For these cases, we will necessarily only approximate the desired distinction. We have interpreted untagged PPs as arguments of the verb. The motivation for this choice comes both from an overall observation of sentences and from the documentation,

-CLR	dative object if dative shift not possible (e.g., *donate*); phrasal verbs; predication adjuncts
-DTV	dative object if dative shift possible (e.g., *give*)
-BNF	benefactive (dative object of *for*)
-PRD	non VP predicates
-PUT	locative complement of *put*
-DIR	direction and trajectory
-LOC	location
-MNR	manner
-PRP	purpose and reason
-TMP	temporal phrases

Figure 2
Grammatical function and semantic tags that involve PP constituents in the Penn Treebank.

in which it is stated that "NPs and Ss which are clearly arguments of the verb are unmarked by any tag" (Marcus et al. 1994, page 4), and that "Direct Object NPs and Indirect Object NPs are all untagged" (Bies et al. 1995, page 12). Although the case of PP constituents is not specifically addressed, we have interpreted these statements as supporting evidence for our choice.

The tag -CLR stands for "closely related," and its meaning varies, depending on the element it is attached to. It indicates argument status when it labels the dative object of ditransitive verbs that cannot undergo dative shift, such as in *donate money to the museum*, and in phrasal verbs, such as *pay for the horse*. It indicates adjunct status when it labels a predication adjunct as defined by Quirk et al. (1985). We interpret the tag -CLR as an argument tag in order not to lose the few cases for which the differentiation is certain: the ditransitive verbs and some phrasal verbs. This choice apparently misclassifies predication adjuncts as arguments. However, for some cases, such as obligatory predication adjuncts, an argument status might in fact be more appropriate than an adjunct status. According to Quirk et al. (1985, Sections 8.27–35, 15.22, pages 16–48), there are three types of adjuncts, differentiated by the degree of "centrality" they have in the sentence. They can be classified into predication adjuncts and sentence adjuncts. Predication adjuncts can be obligatory or optional. Obligatory predication adjuncts resemble objects as they are obligatory in the sentence and they have a relatively fixed position, as in *He lived in Chicago.* Optional predication adjuncts are similarly central in the sentence but are not obligatory, as in *He kissed his mother on the cheek.* Sentence adjuncts, on the contrary, have a lower degree of centrality in the sentence, as in *He kissed his mother on the platform.* As a conclusion, obligatory predication adjuncts as described in Quirk et al. (1985) could be interpreted as arguments, as they are required to interpret the verb (the interpretation of *lived* in *He lived* differs from the one in *He lived in Chicago*).

To recapitulate, we have labeled our examples as follows:

- Adjuncts: All PPs tagged with a semantic tag (DIR, LOC, MNR, PRP, TMP) are adjuncts.
- Arguments: All untagged PPs or PPs tagged with CLR, EXT, PUT, DTV, BNF, or PRD are arguments.

Validating the Target Attribute and Creating a Gold Standard. The overall mapping of Penn Treebank function labels onto the argument–adjunct distinction is certainly too coarse, as some function types of PPs can be both arguments or adjuncts, depending on the head they co-occur with. We assessed the overall validity of the mapping as follows. First, for each of the function tags mentioned earlier, we sampled the Penn Treebank (one example from each section) for a total of 22 examples for each tag. Then, we manually inspected the examples to determine if the mapping onto argument or adjunct was correct. On a first inspection, the tags PUT, DTV, and PRD are correctly mapped as argument in the majority of cases, as well as DIR, LOC, TMP, and PRP, which are correctly considered adjuncts. Those samples tagged MNR, CLR, BNF, or untagged show a more mixed behavior, sometimes appearing to label arguments and sometimes adjuncts. For these labels, we used a more elaborate procedure to determine if the example was an argument or an adjunct. We concentrate on PPs attached to the verb, as these cases appear to be more ambiguous.

All the test examples attached to a verb that had a CLR, PP, or MNR label were extracted. We did not investigate BNF as there aren't any in our test file. We constructed test suites for each example by applying to it the typical linguistic diagnostics used to determine argumenthood, along the lines already discussed in Section 2. Five tests were selected, which were found to be the most discriminating in a pilot study over 224 sentences: optionality, ordering, head dependence, extraction with preposition stranding, and extraction with pied-piping, as illustrated in Figure 3. It can be noticed that we were able to use more complex tests than those used by the algorithm; in particular we use extraction tests (Schütze 1995).[6] A native speaker gave binary acceptability judgments over the 1,100 sentences thus generated. The acceptability judgments were assigned to the sentences over the course of several days. Once the judgments to each quintuple of sentences were collected, they were combined into a single binary-valued target feature for each sentence by the first author. The decision was based on the relative importance and reliability of the tests according to the linguistic literature (Schütze 1995), as follows: *If the optionality test is negative then the example is an argument, else if extraction can take place then the example is an argument, else the majority label according to the outcome of the grammaticality judgments is assigned.* In a few cases, the judgment was overidden if the negative outcome of the tests clearly derived from the fact that the V + PP was an idiom rather than an adjunct: for example, *make fools of themselves, have a ring to it,* and *live up to.*

In the end, we find the following correlations between the automatic labels and those assigned based on the accurate collection of native speaker judgments. Among the PPs that do not have a function label, 65 are adjuncts and 42 are arguments, according to the manual annotation procedure. The label PP-CLR corresponds to 18 adjuncts and 159 arguments, according to our manual annotation procedure, whereas the label PP-MNR corresponds to 5 adjuncts and 1 argument. Clearly, the assignments of CLR PPs to arguments and MNR PPs to adjunct are confirmed. Prepositional phrases without any functional labels, on the other hand, are much more evenly divided between argument and adjunct, as we suspected, given the heterogeneous nature of the label (the label PP

[6] Some of the extracted sentences had to be simplified so that the verb and prepositional phrases were in a main clause. For example *buy shares from sellers*, which is generated from the sentence *On days like Friday, that means they must buy shares from sellers when no one else is willing to* becomes *They must buy shares from sellers*. In some cases the sentences had to be further simplified to allow extraction tests to apply, which would be violated for reasons unrelated to the argument-adjunct distinction in the PP, such as negation, or complex NP islands. For example, *Hill democrats are particularly angry over Mr. Bush's claim that the capital-gains cut was part of April's budget accord and his insistence on combining it with the deficit-reduction legislation* yields *Mr. Bush combines capital gains cut with the deficit-reduction legislation*, which gives rise to the following extraction examples: *What do you wonder whether Mr. Bush combines capital gains cut with? With what do you wonder whether Mr. Bush combines capital gains cut?*

Americans will learn more about making products [for the Soviets].

optionality	Americans will learn more about making products.
order	Americans will learn more about making products these coming years for the Soviets.
head dependence	Americans will learn more about making/selling/reading products for the Soviets.
extraction	Who do you wonder whether Americans will learn more about making products for?
	For who(m) do you wonder whether Americans will learn more about making products?

Figure 3
Example sentence and related list of tests and test sentences.

is assigned not only to those cases that are clear arguments, but also to those cases for which a label cannot be decided). Moreover, if the hand-annotated label is reliable, it indicates that untagged PPs are somewhat more likely to be adjuncts. Our initial mapping was incorrect. In retrospect, we must conclude that other researchers had applied what appears to be the correct mapping for this data set (Buchholz 1999). We do not, however, modify the label of our training set, as that would be methodologically incorrect. We have relabeled the test set and are therefore bound to ignore any further knowledge we have gathered in relabeling, as that would amount to tailoring our training set to our test set. The consequence of this difference between our label and what we found to be true for the gold standard is that all results tested on the manually labeled test set will have to be interpreted as lower bounds of the performance to be expected on a consistently labeled data set. Besides validating the automatically annotated test set, the manually annotated test set serves as a gold standard. Performance measures on this set will support comparison across methodologies.

Therefore, we conclude that the mapping we have assumed is coherent with the judgments of a native speaker, although the agreement is not perfect. PPs without a function tag are an exception. Thus, the automatic mapping we have defined will provide the value of the target feature in the experiments that we illustrate in the two following sections. When appropriate we report two performance measures, one for the automatic label and one for the partly manual labels. We will also report some comparative results on a test set that does not contain the noisy PPs without function labels.

4. Distinguishing Arguments from Adjuncts

Having collected the necessary data and established the value of the target attribute for each example, we can now perform experiments to test several different hypotheses concerning the learning of the argument-adjunct distinction. First of all we need to show that the distinction under discussion can be learned to a good degree. Furthermore, we investigate the properties that, singly or in combination, lead to an improvement in learning. We are particularly interested in comparing the baseline to a simple model, where learning is done using only lexical heads. In this case, we investigate the relevance of the simple words in detecting the difference between arguments and adjuncts. We also verify the usefulness of knowing lexical classes on the accuracy of learning. Finally, we show that the diagnostic features and the lexical classes developed above bring more information to the classification than what can be gathered by simple lexical heads. We summarize these expectations below, where we repeat and elaborate Hypothesis 1 formulated in the introduction.

- Hypothesis 1: The argument-adjunct distinction can be performed based on information collected from an annotated corpus.
- Hypothesis 1′: The argument-adjunct distinction can be improved by lexical classes and linguistic diagnostic features, (i) over a simple baseline, but also (ii) over a model using lexical features.

Demonstration of these hypotheses requires showing that the distinction can be learned from corpus evidence, even with a simple method (performance is better than chance). Hypothesis 1′ imposes the more stringent condition that we can considerably improve a simple learner by using more linguistically informed statistics (performance is better than the simple method).

4.1 The Input Data and the Classifier

In order to test the argument and adjunct attachment problem independently of whether the PP is attached to a verb or to a noun, we create two sets of input data, one for verb attachment and one for noun attachment.

Each input vector for verb attachment contains training features comprising the four lexical heads and their WordNet classes (v, $n1$, p, $n2$, v_{cl}, $n1_{\mathrm{cl}}$, and $n2_{\mathrm{cl}}$), all the different variants of the implementation of the diagnostics for the argument-adjunct distinction concerning PPs attached to a verb, and one binary target feature, indicating the type of attachment, whether argument or adjunct. More specifically, the features implementing the diagnostics for verbs consist of the variants of the measures of head dependence (*hdepv1,hdepv2*), the variants of the measure of optionality (*opt1, opt2, opt3*), and the measure of iterativity (*iterv*).

Each input vector for noun attachment contains 14 training features. They comprise the four lexical heads and their WordNet classes, as above (v, $n1$, p, $n2$, v_{cl}, $n1_{\mathrm{cl}}$, and $n2_{\mathrm{cl}}$), all the different variants of the implementation of the diagnostics for PPs attached to a noun, and one binary target feature, indicating the type of attachment. The features implementing the diagnostics for nouns are: the variants of the measures of head dependence (*hdepn1, hdepn2, hdepn3*), the measures of iterativity (*itern*), and the measures for copular paraphrase and deverbal noun, respectively (*para, deverb*).

For both types of experiments—distinction of arguments from adjuncts of PPs attached to the noun or PPs attached to the verb—we use the C5.0 Decision Tree Induction Algorithm (Quinlan 1993) and Support Vector Machines (LIBSVM), version 2.71 (Chang and Lin 2001).

4.2 Results on V Attachment Cases

We have run experiments on the automatically labeled training and test sets with very many different feature combinations. A summary of the most interesting patterns of results are indicated in Tables 1 and 2. Significance of results is tested using a McNemar test.

Contribution of Lexical Items and Their Classes. In this set of experiments we use only lexical features or features that encode the lexical semantic classes of the open class

Table 1
Accuracy of the argument–adjunct distinction for VP-attached PPs, using combinations of lexical features. The training and test sets are automatically annotated.

Feature used	Auto accuracy (%)
1. Chance (args)	55.8
2. Prep (baseline)	67.9
3. Lexical features (verb, prep, n2)	67.9
4. $\mathrm{v_{cl}}$, prep	73.8
5. Prep, $\mathrm{n2_{cl}}$	71.9
6. Lexical features and classes (v, $\mathrm{v_{cl}}$, p, n2, $\mathrm{n2_{cl}}$)	68.2
7. Only classes ($\mathrm{v_{cl}}$, p, $\mathrm{n2_{cl}}$)	75.0
8. Only verb classes (v_{cl}, p, n2)	73.1
9. Only noun classes (v, p, $n2_{\mathrm{cl}}$)	70.3
10. All features	68.2

words in question. Lines 2 to 5 of Table 1 report better classification values than line 1. These results indicate that it is possible, at different degrees of accuracy, to leverage information encoded in lexical items to infer the semantic distinction between arguments and adjuncts without explicit deep semantic knowledge. One interesting fact (lines 2 and 3) is that the preposition is very informative, as informative as all the lexical items together. An analysis of the distribution of arguments and adjuncts by preposition indicates that whereas most prepositions are ambiguous, they have a strong preference for either arguments or adjuncts. Only a few equibiased prepositions, such as *for*, exist. An expected result (lines 4 and 5) is that the PP-internal noun class is useful, in combination with the preposition, as well as the combination of verb class and preposition (the difference is marginally significant). We find the best result (line 7) in the experiment that uses class combinations (difference from baseline $p < .001$). We see that classes of open class items, nouns, and verbs associated with the closed class item preposition give the best performance. Line 7 is considerably better than line 6 ($p < .001$), indicating that the actual presence of individual lexical items is disruptive. This indicates that regularities in the distinction of arguments from adjuncts is indeed a class phenomenon and not an item-specific phenomenon. This is expected, according to current linguistic theories, such as the one proposed by Levin (1993).

This analysis is confirmed by observing which are the most discriminative features, those that are closest to the root of the decision tree. The topmost feature of the best result (line 7) is *preposition*, followed by the class of the verb and the class of the PP internal noun. Predictably, the class of the object noun phrase is not used because it is not very informative. The same features constitute the tree yielding the results of line 6. However, the presence of lexical items makes a difference to the tree that is built, which is much more compact, but in the end less accurate.

Combinations of All Features. Table 2 reports the accuracy in the argument-adjunct distinction of experiments that use only the most useful lexical and class features, the preposition and the verb class, and the diagnostic-based features, using combinations of diagnostic features. The combinations of features shown are those that yielded the best results over a development set of tuples extracted from Section 24 of the Penn Treebank. The results reported are calculated over a test corresponding to the tuples

Table 2
Best results using preposition and combination of diagnostic-based features, in different variants. The training and test sets are automatically annotated.

Feature used	Auto accuracy (%)
1. v_{cl}, prep, hdepv1, hdepv2, opt1, opt2, opt3, iterv	79.3
2. v_{cl}, prep, hdepv1, hdepv2, opt1, opt3, iterv	79.8
3. v_{cl}, prep, hdepv1, hdepv2, opt2, opt3, iterv	79.0
4. v_{cl}, prep, hdepv1, opt2, opt3, iterv	78.2

in Section 23 of the Penn Treebank. The best combination, line 2, yields a 37% reduction of the error rate over the baseline. All the differences in performance among these configurations are significant. What all these combinations have in common is that they are combinations of three levels of granularity, mixing lexical information, class information, and higher level syntactic-semantic information, encoded indirectly in the diagnostics. All the combinations that are not shown here have lower accuracies.

Table 3 shows the confusion matrix for the combination of features with the best accuracy listed above. These figures yield a precision and recall for arguments of 80% and 85%, respectively (*F* measure = 82%); and a precision and recall for adjuncts of 80% and 73%, respectively (*F* measure = 76%). Clearly, although both kinds of PPs are well identified, arguments are better identified than adjuncts, an observation already made by several other authors, especially Hindle and Rooth (1993) in their detailed discussion of the errors in a noun or verb PP-attachment task. In particular, we notice that more adjuncts are misclassified as arguments than vice versa.

The results of these experiments confirm that corpus information is conducive to learning the distinction under discussion without explicitly represented complex semantic knowledge. They also confirm that this distinction is essentially a word class phenomenon—and not an individual lexical-item phenomenon—as would be expected under current theories of the syntax–semantics interface. Finally, the combination of lexical items, classes, and linguistic diagnostics yields the best results. This indicates that using features of different levels of granularity is beneficial, probably because the algorithm has the option of using more specific information when reliable, while abstracting to coarser-grained information when lexical features suffer from sparse data. This interpretation of the results is supported by observing which features are at the top of the tree. Interestingly, here the topmost feature is *head dependence* (the lexical variant, hdepv1), on one side of which we find *preposition* as the second most discriminative feature, followed by *head dependence* (hdepv2) again, and *optionality* (class variants). On

Table 3
Confusion matrix of the best classification of PPs attached to the verb. Training and test set established automatically.

		Assigned classes		
		Arguments	Adjuncts	Total
Actual classes	Arguments	300	51	351
	adjuncts	76	202	278
	Total	376	253	629

the other side of the tree, we find *preposition* as the second most informative feature and *verb class* as the third most discriminative feature.

Results on Partly Manually Labeled Set. Tables 4 and 5 report the results obtained by training the classifier on the automatically labeled training set and testing on the manually labeled test set. They illustrate the effect of training the decision tree classifier on a training set that has different properties from the test set. This experiment provides a lower bound of performance across different samples and shows which are the features with the greatest generalization ability. We can draw several conclusions. First, the lexical features do better than chance, but do not do better than the baseline established by using only the preposition as a feature (lines 1, 2, and 3 of Table 4). Secondly, classes do better than the baseline (line 7 of Table 4) and so do the diagnostic features (Table 5). Since we are using a training and a test set with different properties, these results indicate that classes and diagnostics capture a level of generality that the lexical features do not have and will be more useful across domains and corpora. Finally, the rank of performance for different feature combinations holds across training and testing methods, whether established automatically or manually, as can be confirmed by a comparison of Tables 1 and 4 and also 2 and 5. The difference in performance with diagnostics (line 3 of Table 5) and without, using only classes (line 7 of Table 4), is only marginally significant, indicating that diagnostics are not useless.

Table 4
Accuracy of the argument–adjunct distinction for VP-attached PPs, using combinations of lexical features. The training set is automatically annotated while the test set is in part annotated by hand.

Feature used	Manual accuracy (%)
1. Chance (args)	37.0
2. Prep (baseline)	61.2
3. Lexical features (verb, prep, n2)	61.2
4. v_{cl}, prep	67.1
5. Prep, $n2_{cl}$	66.8
6. Lexical features and classes (v, v_{cl}, p, n2, $n2_{cl}$)	62.8
7. Only classes (v_{cl}, p, $n2_{cl}$)	70.0
8. Only verb classes (v_{cl}, p, n2)	67.7
9. Only noun classes (v, p, $n2_{cl}$)	64.9
10. All features	66.0

Table 5
Best results using prepositions and combination of diagnostic-based features in different variants. The training set is automatically annotated, whereas the test set is in part annotated by hand.

Feature used	Manual accuracy (%)
1. Baseline (prep)	67.3
2. v_{cl}, prep, hdepv1, hdepv2, opt1, opt2, opt3, iter	67.2
3. v_{cl}, prep, hdepv1, hdepv2, opt1, opt3, iter	69.0
4. v_{cl}, prep, hdepv1, hdepv2, opt2, opt3, iter	67.6
5. v_{cl}, prep, hdepv1, opt2, opt3, iter	65.7

Results on Test Set without Bare PPs. The biggest discrepancy in validating the automatic labeling was found for PPs without functional tags. The automatic labeling had classified bare PPs as argument but the manual gold standard assigns more than half of them to the adjunct class. They are therefore a source of noise in establishing reliable results. If we remove these PPs from the training and test set, results improve and become almost identical across the manually and automatically labeled sets, as illustrated in Table 6.[7]

4.3 Results on N Attachment Cases

Experiments on learning the distinction between argument PPs and adjunct PPs attached to a noun show a different pattern, probably due to a ceiling effect. The experiments reported below are performed on examples of prepositional phrases whose preposition is not *of*. The reason to exclude the preposition *of* is that it is 99.8% of the time attached as an argument. Moreover, it accounts for approximately half of the cases of NP attachment. Results including this preposition would therefore be overly optimistic and not be representative of the performance of the algorithm in general. The size of the resulting corpus—without *of* prepositional phrases—is 3,364 tuples. The results illustrated in Tables 7 and 8 show that head dependence is the only feature that improves numerically the performance above the already very high baseline that can be obtained by using only the feature *preposition*. This difference is not statistically significant, indicating that neither class nor diagnostics add useful information.

4.4 Results Using a Different Learning Algorithm

In the previous sections, we have shown that different combinations of features yield significantly different performances. We would like to investigate, at least on a first approximation, if these results hold across different learning algorithms. To test the stability of the results, we compare the main performances obtained with a decision tree to those obtained with a different learning algorithm. In recent years, a lot of attention has been paid to large margin classifiers, in particular support vector machines (SVMs) (Vapnik 1995). They have been shown to perform quite well in many NLP tasks. The fact that they search for a large separating margin between classes makes them less prone to overfitting. We expect them, then, to perform well on the task trained on automatically labeled data and tested on manually labeled data, where a great ability to generalize is needed. Despite being very powerful, however, SVMs are complex algorithms, often opaque in their output. They are harder to interpret than decision trees, where the position of the features in the tree is a clear indication of their relevance for the classification. Finally, SVMs take a longer time to train than decision trees. For these reasons they constitute an interestingly different learning technique from decision trees, but not a substitute for them if clear interpretation of the induced learner is needed and many experiments need to be run.

All the experiments reported below were performed with the LIBSVM package (Chang and Lin 2001, version 2.71). The SVM parameters were set by a grid search on the training set, by 10-fold cross-validation. Given the longer training times, we perform only a few experiments. For V attachment, the baseline using only the preposition

[7] The features indicated here as best are the ones used in the other experiments and kept for the sake of comparison. But, in fact, a different feature combination found by accident gives a result of 79.3% accuracy in classifying the automatically labeled set.

Table 6
Best results using preposition and combination of diagnostic-based features, in different variants, taking out PP examples.

Feature used	Manual accuracy (%)	Auto accuracy (%)
1. Chance baseline (args)	67.3	37.0
2. Prep baseline	64.4	64.6
3. Classes	74.5	74.7
4. Best features combination	76.1	77.0

Table 7
Baselines and performances using lexical heads and classes for N-attached PPs.

Features used	Accuracy (%)
chance (arg)	58.3
p (baseline)	93.3
n1, p, n2	93.3
$n1_{cl}$, p, $n2_{cl}$	93.3
n1, $n1_{cl}$, p, n2, $n2_{cl}$	93.3

reaches an accuracy of 62.2% on the manually labeled test set, whereas performance using the same best combination of features as the decision tree reaches 70.6% accuracy. There is, then, a little improvement over the 69% of the decision tree learner, as expected. The performance on the N-attached cases, on the other hand, is surprisingly poor, with a low 43% accuracy on testing data. This result is probably due to overfitting, since the best accuracy on the training set is around 95%.

4.5 Conclusions

The results reported in this section show that the argument-adjunct distinction can be learned based on information collected from an annotated corpus with good accuracy. For verb attachment, they show in particular that using lexical features yields better performance than the baseline, especially when we use lexical classes. For automatically labeled data, diagnostics based on linguistic theory improve the performance even further. Thus, the hypotheses we were testing with these experiments are confirmed.

The reported results are good enough to be practically useful. In particular, the distinction between arguments and adjuncts attached to nouns is probably performed as well as possible with an automatic method, even by simply using prepositions as features. For the attachment to verbs, known to be more difficult, more room for

Table 8
Performances using some combinations of features for N-attached PPs.

Features used	Accuracy (%)
$n1_{cl}$, prep, $n2_{cl}$, hdepn1, hdepn2, hdepn3, itern, para, deverb	93.9
prep, hdepn1, hdepn2, hdepn3, itern, para, deverb	93.9
prep, hdepn1	94.1

improvement exists, especially in the recovery of adjuncts. The comparison of decision trees to SVMs does not appear to indicate that one learning algorithm is consistently better than the other.

5. Hypothesis 2: PP Attachment Disambiguation

Once we have established the fact that arguments and adjuncts can be learned from a corpus with reasonable accuracy using cues correlated to linguistic diagnostics, we are ready to investigate how this distinction can be integrated in the disambiguation of ambiguously attached PPs, the PP attachment problem as usually defined.

The first question that must be asked is whether the distinction between arguments and adjuncts is so highly correlated with the attachment site of the ambiguous PP to be almost entirely derivative. For example, the PTB annotators have annotated all noun attachments as adjuncts and all verb attachments as arguments. If this were the correct representation of the linguistic facts, having established an independent procedure to discriminate argument from adjuncts PPs would be of little value in the disambiguation problem. In fact, there is no theoretical reason to think that the notion of argument is closely correlated to the choice of attachment site of a PP, given that both verb and noun attached PPs can have either an argument or an adjunct function. It might be, however, that some distributional differences that are lexically related, or simply nonlinguistic, exist and that they can be exploited in an automatic learning process.

We can test the independence of the distribution of arguments and adjuncts from the distribution of noun or verb attachment with a χ^2 test. The test tells us that the two distributions are not independent ($p < .001$). It remains, however, to be established if the dependence of the two distributions is sufficiently strong to improve learning of one of the two classifications, if the other is known. This question can be investigated empirically by augmenting the training features for a learning algorithm that solves the usual binary attachment problem with the diagnostic features for argumenthood. If this augmentation results in an improvement in the accuracy of the PP attachment, then we can say that the notion of argument is, at least in part, related to the attachment site of a PP. If no improvement is found, then this confirms that the argument status of a PP must be established independently. Conversely, we can augment the input to the classification into argument and adjuncts with information related to the attachment site to reach analogous conclusions about the distinction between arguments and adjuncts.

Corpora and Materials. The data for the experiments illustrated below are drawn from the same corpora as those used in the previous experiments. In particular, recall that the corpus from which we draw the statistics is different from the corpus from which we draw the training examples and also from the testing corpus. In both sets of experiments described below, we restrict our observation to those examples in the corpus that are ambiguous between the two attachment sites, as is usual in studies of PP attachment. The values of the learning features were calculated on all the instances in the statistics corpus, so in practice we use both the unambiguous cases and ambiguous cases in the estimation of the features of the ambiguous cases.

The Input Data. Each input vector represents an instance of an ambiguous PP attachment, which could be both noun or verb attached, either as an argument or as an adjunct. Each vector contains 20 training features. They comprise the four lexical heads, their

WordNet classes (*v, n*1*, p, n*2*,* v_{cl}*,* $n1_{cl}$, and $n2_{cl}$), and all the different variants of the implementation of the diagnostics. Finally, depending on the experiments reported below we use either a two-valued target feature (N or V) or a four-valued target feature (Narg, Nadj, Varg, Vadj), indicating the type of attachment. More specifically, the features implementing the diagnostics are the variants of the measures of head dependence for the PPs attached to verbs and nouns, respectively (*hdepv1, hdepv2,* and *hdepn1, hdepn2, hdepn3*); the variants of the measure of optionality (*opt1, opt2, opt3*); the measures of iterativity for verb-attached and noun-attached PPs, respectively (*iterv, itern*); and finally the measures for copular paraphrase and deverbal noun, respectively (*para, deverb*).

We use the C5.0 Decision Tree Induction Algorithm (Quinlan 1993), and the implementation of SVMs provided in version 2.71 of LIBSVM (Chang and Lin 2001).

5.1 Relationship of Noun–Verb Attachment Disambiguation to the Argument-Adjunct Distinction and Vice Versa

Here we report on results for the task of disambiguating noun or verb attachment first, using, among other input features, those that have been established to make the argument-adjunct distinction. The same corpora described above were used, with a two-valued target (N or V). We report results for two sets of experiments. One set of experiments takes all examples into account. In another set of experiments, examples containing the preposition *of* were not considered, as this preposition is extremely frequent (it adds up to almost half of the noun attachment cases) and it is almost always attached to a noun as argument. It has therefore a very peculiar behavior. The best combination of features reported below was established using Section 24 of the Penn Treebank; all the tests reported here are in Section 23.

Table 9 reports the disambiguation accuracy of the comparative experiments performed. The first line reports the baseline accuracy for the task, calculated by performing the classification using only the feature *preposition*. The best result is obtained by a combination of features in which lexical classes act as the predominant learning feature, either in combination with the lexical items or alone (line 2 with *of*, line 3 without *of*). We note, however, that the diagnostic features that are included in the best diagnostic feature combination are those based in part on individual words and not those based entirely on classes. Most importantly, those diagnostics that are meant to directly indicate the argument or adjunct status of a PP do not help in the resolution of PP attachment, as expected.

Table 9
Percent accuracy using combinations of features for two-way attachment disambiguation. Best combinations for experiments with *of* is (opt1, hdepv1, hdepn1, para) and for experiments without *of* is (opt1, opt2, hdepv1, hdepv2, hdepn1, hdepn2, para).

Features used	Accuracy with *of* (%)	Accuracy without *of* (%)
1. Prep (baseline)	70.9	59.5
2. Prep + classes	78.1	71.3
3. Only classes	70.2	72.3
4. Only all diagnostics	75.9	64.5
5. Prep + all diagnostics	77.1	68.2
6. Prep + best feature combination	75.8	67.8
7. Prep + classes + best feature combination	76.6	67.5

We conclude, then, that the notion of argument and adjunct is only partially correlated to the classification of PPs into those that attach to the noun and those that attach to the verb. Clearly, diagnostics are not related to the attachment site, but lexical classes appear to be. On the one hand, this result indicates that the notion of argument is not entirely derivative from the attachment site. On the other hand, it shows that some features developed to distinguish arguments from adjuncts could improve the disambiguation of the attachment site.

The same conclusion is confirmed by a simpler, much more direct experiment, where the classification into noun or verb attachment is performed with a single input attribute. This attribute is the feature indicating if the example is an argument or an adjunct, and it is calculated by a binary decision tree classifier using the best feature combination on the argument-adjunct discrimination task. In this case too the classification accuracy is a little (2.5%) better than chance baseline.

The converse experiment does not reveal any correlations, confirming that the interdependence between the two factors is weak. The attachment status of the ambiguous PP, whether noun or verb attached, is input among other features, to determine whether the PP is an argument or an adjunct. Results are shown in Table 10, where the attachment feature is called *NVstatus*. Lines 1 and 2 show that NVstatus is a better baseline than chance. Lines 3 and 4 indicate that the feature *preposition* offers a good baseline over which NVstatus improves only if the preposition *of* is included, as expected. Lines 5 and 6 and lines 7 and 8 show that adding NVstatus to the other features does not improve performance. As previously, the lexical classes are the best performing features.

The same conclusion is reached by a simple direct experiment where we classify PPs into arguments and adjuncts using as only input feature the output of a classifier between noun or verb attachment. This attachment classifier is trained on the best feature combination, preposition, and word classes. We find that the attachment status has no effect on the accuracy of the classification, as the feature is not used.

Overall, these results indicate that there is a small interdependence between the two classification problems and therefore weakly support a view of the PP disambiguation problem where both the site of the attachment and the argument or adjunct function of the PP are disambiguated together in one step. In the next section, we explore this hypothesis, while also investigating which feature combinations give the best results in a four-way PP classification, where PPs are classified as noun arguments, noun adjuncts, verb arguments, and verb adjuncts.

Table 10
Percent accuracy using combinations of features for argument adjunct classification, including NV as input feature. Best features combination is (opt1, opt3, hdepv1, hdepv2, hdepn1).

Features used	Accuracy with *of* (%)	Accuracy without *of* (%)
1. Chance (args)	69.6	55.9
2. NVstatus	73.0	62.3
3. Prep (baseline)	81.6	82.0
4. NVstatus + prep	87.2	81.5
5. Prep + classes	89.2	84.6
6. Prep + classes + NVstatus	89.0	84.1
7. Prep + classes + best features + NVstatus	88.2	82.9
8. Prep + classes + best features	88.2	83.6

5.2 One- and Two-step Four-way Classification

Having shown that argumenthood of a PP is not entirely derivative of its attachment site, but that the two tasks are weakly correlated, the task of PP attachment itself is reformulated as a four-way classification, yielding a finer-grained and more informative classification of PP types.

As discussed in the introduction, those applications for which the resolution of the PP attachment ambiguity is important often need to know more about the PP than its attachment site, in particular one might need to know whether the PP is an argument or not. For example, the output of a parser might be used to determine the kernel of a sentence—the predicate with its arguments—for further text processing, for translation, or for constructing a lexicon. We redefine therefore the problem of PP attachment as a four-way classification problem. We investigate here what features are best predictors of this classification. Again, we report results for two sets of experiments. One set of experiments takes all examples into account, whereas the other excludes all examples including the preposition *of*. As usual, the best combination of features reported below was established using Section 24; all the tests reported here are in Section 23.

To classify PPs into four classes, we have two options: We can construct a single four-class classifier or we can build a sequence of binary classifiers. The discrimination between noun and verb attachment can be performed first, and then further refined into attachment as argument or adjunct, performing the four-way classification in two steps. The two-step approach would be the natural way of extending current PP attachment disambiguation methods to the more specific four-way attachment we propose here. However, based on the previous experiments, which showed a limited amount of dependence between the two tasks, previous work on a similar data set (Merlo 2003), and general wisdom in machine learning, there is reason to believe that it is better to solve the four-way classification problem directly rather than first solving a more general problem and then specializing the classification.

To test these expectations, we performed both kinds of experiments—a direct four-way classification experiment and a two-step classification experiment—to investigate which of the two methods is better. The direct four-way classification uses the attributes described above to build a single classifier. For comparability, we created a two-step experimental setup as follows. We created three binary classifiers. The first one performs the noun–verb attachment classification. Its learning features comprise the four lexical heads and their WordNet classes. We also train two classifiers that learn to distinguish arguments from adjuncts. One classifier is trained only on verb-attachment exemplars and uses only the best verb-attachment-related features. The third classifier is trained only on noun-attachment exemplars and utilizes only the best noun-attachment-related features. The test data is first given to the noun–verb attachment classifier. Then, the test examples classified as verbs are given to the verb argument-adjunct classifier, and the test examples classified as nouns are given to the noun argument-adjunct classifier. Thus, this cascade of classifiers performs the same task as the four-way classifier, but it does so in two passes.

Table 11 shows that overall the one-step classification is better than the two-step classification, confirming the intuition that the two labeling problems should be solved at the same time.[8] However, if we break down the performance, we see that recall of

[8] The difference between the two results is significant ($p < .05$) according to the randomized test described in Yeh (2000).

Table 11
Percent precision, recall, and *F* score for the best two-step and one-step four-way classification of PPs, including and not including the preposition *of*.

	Two-step + of			One-step + of		
	Prec	Rec	*F*	Prec	Rec	*F*
V-arg	37.5	45.6	41.2	42.2	29.3	34.6
V-adj	56.2	52.2	54.1	59.6	60.2	59.9
N-arg	83.0	83.5	83.2	81.3	91.3	86.0
N-adj	71.2	57.5	63.6	69.5	56.2	62.1
Accuracy		68.9			72.0	

	Two-step − of			One-step − of		
	Prec	Rec	*F*	Prec	Rec	*F*
V-arg	41.3	50.0	45.3	42.2	31.4	36.0
V-adj	52.8	41.6	46.5	59.6	60.2	59.9
N-arg	67.3	70.0	68.6	65.4	80.7	71.9
N-adj	60.3	60.3	60.3	69.5	56.2	62.1
Accuracy		56.6			60.9	

V-arg is lower in the one-step procedure than in the two-step procedure, in both cases, and that the overall performance for V-arg is worse in the one-step procedure. This might indicate that which procedure to use depends on whether precision or recall or overall performance is most important for the application at hand.

Table 12 reports the confusion matrix of the classification that reaches the best performance without the preposition *of*, which corresponds to the lower right panel of Table 11. It can be observed that performances are reasonably good for verb adjuncts, and noun arguments and adjuncts, but they are quite poor for the classification of prepositional phrases that are arguments of the verb. It is not clear why verb arguments are so badly classified. We tested the hypothesis that this result is a side effect of the mapping we have defined from the Penn Treebank label to the label we use in this classifier, arguments or adjuncts. Recall that untagged PPs have been mapped onto the argument label, but these are highly inconsistent labels, as we have seen in the manual validation of the target attribute in Section 3. Then, verb arguments might be represented by noisier training examples. This hypothesis is not confirmed. In a little experiment where the training data did not contain untagged verb-attached PPs, the overall performance in identifying the verb argument class did not improve. An improvement in precision was counteracted by a loss in recall, yielding slightly worse *F* measures. Another observation related to verb arguments can be drawn by comparing the experiments reported in Section 4 to the experiments reported in the current section. This comparison shows that the low performance in classifying verb arguments does not arise because of an inability to distinguish verb arguments from verb adjuncts. Rather, it is the interaction of making this distinction and disambiguating the attachment site as a single classification task that creates the problem. This is also confirmed by the considerable number of cases of noun arguments and verb arguments that are incorrectly classified, as shown in Table 12. Clearly, further study of the properties of verb arguments is needed.

Table 12
Confusion matrix of the best one-step four-way classification of PPs without the preposition *of*.

		Assigned classes				
		V-arg	V-adj	N-arg	N-adj	Total
Actual classes	V-arg	27	17	39	3	86
	V-adj	15	68	18	12	113
	N-arg	19	7	121	3	150
	N-adj	3	22	7	41	73
	Total	64	144	185	59	422

Overall, it is interesting to notice that solving the four-way task causes only a little degradation to the accuracy of the original disambiguation between attachment to the noun or to the verb. On this data set, the accuracy of disambiguating the attachment site is of 83.6% (without PPs containing *of*). Accuracy decreases a little to 82.7% if the binary attachment disambiguation result is calculated on the output of the four-way task. This little degradation is to be expected as the four-way task is more difficult. The accuracy of the four-way task on the simple noun or verb binary attachment distinction seems, however, acceptable, if one considers that a finer-grained discrimination is provided.

Table 13 reports the classification accuracy of a set of comparative experiments. Here again, the first line reports the baseline accuracy for the task, calculated by performing the classification using only the feature *preposition*. We notice that in both columns the best results are obtained by the same combination of features that includes some lexical features, some classes, and some diagnostic features. This shows that the distinction between arguments and adjuncts is not exclusively a syntactic phenomenon and lexical variability plays an important role. Similarly to the previous set of experiments, we note that the diagnostic features that are included in the best feature combination are those based, at least in part, on individual words, and not those based entirely on classes. The importance of lexical classes, however, is confirmed by the fact that the best result is only marginally better than the second best result, in which lexical classes act as the predominant learning feature, either in combination with the lexical items or alone (line 3 with *of*, line 2 without *of*). We can conclude from these observations that using various features defined at different levels of granularity allows the learner to better use lexical information when available, and to use more abstract

Table 13
Percent accuracy using combinations of features for a one-step four-way classification of PPs. Best combination = (v_{cl}, $n1_{cl}$, p, opt1, opt2, hdepv1, hdepv2, hdepn1, para).

Features	Accuracy (%) with *of*	Without *of* (%)
1. Prep (baseline)	64.2	49.5
2. Prep + classes	68.9	60.2
3. Only classes	71.5	49.3
4. All features	68.9	54.5
5. Only diagnostics	67.4	54.3
6. Best combination	72.0	60.9

levels of generalization when finer-grained information is not available. Across the two tasks (illustrated in Tables 9 and 13) we notice that the best diagnostics features are almost always the same. Variants change, but the kinds of diagnostics that are useful remain stable. This probably indicates that some diagnostics are reliably estimated, whereas others are not, and cannot be used fruitfully.

5.3 Results Using Support Vector Machines

As mentioned above, SVMs have yielded very good results in many important applications in NLP. It is reasonable to wonder if we can improve the results for the four-way classification, and especially the less than satisfactory performance on verb arguments, using this learning algorithm. Table 14 shows the results to be compared to those in the right-hand panel of Table 11.

If we consider the *F*-measures of this table and the right-hand panel of Table 11, the most striking difference is that V-arguments, although still the worst cell in the table, have improved by almost 20% (36.0% vs. 55.9%) in performance for the experiments without *of*. Notice that verb arguments are now better classified than in the two-step method. Also, the overall accuracy is significantly improved by several percentage points, especially for the condition without the preposition *of* ($p < .02$).

5.4 Conclusion

In this section we have shown that the notion of argument is of limited help in disambiguating the attachment of ambiguous PPs, indicating that the two notions are not strictly related and must be established independently. In a series of four-way classification experiments, we show that the classification performances are reasonably good for verb adjuncts, noun arguments, and noun adjuncts, but they are poor for the classification of prepositional phrases that are arguments of the verb, if decision trees are used. Overall performance and especially identification of verb arguments is improved if support vector machines are used. We also show that better accuracy is achieved by performing the four-way classification in one step. The features that appear to be most effective are lexical classes, thus confirming current linguistic theories that postulate that a given head's argument structure depends on the head's lexical semantics, especially for verbs (Levin 1993).

Table 14
Percent precision, recall, and *F*-score for the best four-way classification of PPs, including and not including the preposition *of* using SVMs.

	One-step + of			One-step − of		
	Prec	Rec	*F*	Prec	Rec	*F*
V-arg	59.5	47.8	53.0	60.0	52.3	55.9
V-adj	63.2	63.7	63.4	62.8	62.8	62.8
N-arg	83.6	93.4	88.2	69.9	85.3	76.9
N-adj	72.5	50.7	59.7	72.5	50.7	59.7
Accuracy		75.9			66.6	

6. Related Work

The resolution of the attachment of ambiguous PPs is one of the staple problems in computational linguistics. It serves as a good testing ground for new methods as it is clearly defined and self-contained. We have argued, however, that it is somewhat oversimplified, because knowing only the attachment site of a PP is of relatively little value in a real application. It would be more useful to know where the PP is attached and with what function. We review below the few pieces of work that have tackled the problem of labeling PPs by function (arguments or adjuncts) as a separate labeling problem. Other pieces of work have asked a similar question in the context of acquiring high-precision subcategorization frames. We review a few of them below.

6.1 On the Automatic Distinction of Arguments and Adjuncts

A few other pieces of work attempt to distinguish PP arguments from adjuncts automatically (Buchholz 1999; Merlo and Leybold 2001; Villavicencio 2002). We extend and modify here the preliminary work reported in Merlo and Leybold (2001) by extending the method to noun attachment, elaborating more learning features, including cases specifically developed for noun attachment, refining all the counting methods, thus validating and extending the approach.

The current work on automatic binary argument-adjunct classifiers appears to compare favorably to the only other study on this topic (Buchholz 1999). Buchholz (1999) reports an accuracy of 77% for the argument-adjunct distinction of PPs performed with a memory-based learning approach, to be compared with our 80% and 94% for verb and noun attachments, respectively. However, the comparison cannot be very direct, as Buchholz considers all types of attachment sites, not just verbs and nouns.

More recently, Villavicencio (2002) has explored the performance of an argument identifier, developed in the framework of a model of child language learning. Villavicencio concentrates on locative PPs proposing that the distinction between obligatory arguments, optional arguments, and adjuncts is made based on two features: a feature derived from a semantically motivated hierarchy of prepositions and predicates and a simple frequency cutoff of 80% of co-occurrence between the verb and the PP that distinguishes obligatory arguments from the other two classes. She evaluates the verbs *put, come,* and *draw* (whose locative arguments belong to the three classes above, respectively). The approach is not directly comparable, as it is not entirely corpus-based (the input to the algorithm is an impoverished logical form), and the evaluation is on a smaller scale than the present work. On a test set of the occurrences of three verbs, which is the same set inspected to develop the learning features, Villavicencio gets perfect performance. These are very promising results, but because they are not calculated on a previously unseen test set, the generability of the approach is not clear. Moreover, Villavicencio applies only one diagnostic test to determine if a PP is an argument or an adjunct, whereas our extensive validation study has shown that several tests are necessary to reach a reliable judgment.

In a study about the automatic acquisition of lexical dependencies for lexicon building, Fabre and Bourigault (2001) discuss the relation of the problem of PP attachment and the notion of argument and adjunct. They correctly notice that approaches such as theirs, inspired by Hindle and Rooth (1993), are based on the assumption that high co-occurrence between words is an indication of a lexical argumenthood relation. As also noticed in Merlo and Leybold (2001), this is not always the case: some adjuncts frequently co-occur with certain heads too. Fabre and Bourigault propose a notion of productivity that strongly resembles our notions of optionality and head dependence to capture the two intuitions about the distribution of arguments and adjuncts. Arguments are strongly selected by the head (the head to complement relation is not productive), whereas adjuncts can be selected by a wide spread of heads (the complement to head selection is highly productive). They propose, but do not test, the hypothesis that this notion might be useful for the general problem of PP attachment. The results in the current article show that this is not the case. In fact, we have argued that there is no real reason to believe that the two notions should be related, other than marginally.

6.2 On the Distinction of Argument from Adjunct PPs for Subcategorization Acquisition

As far as we are aware, this is the first attempt to integrate the notion of argumenthood in a more comprehensive formulation of the problem of disambiguating the attachment of PPs. Hindle and Rooth (1993) mention the interaction between the structural and the semantic factors in the disambiguation of a PP, indicating that verb complements are the most difficult. We confirm their finding that noun arguments are more easily identified, whereas verb complements (either arguments or adjuncts) are more difficult. Other pieces of work address the current problem in the larger perspective of distinguishing arguments from adjuncts for subcategorization acquisition (Korhonen 2002a; Aldezabal et al. 2002).

The goal of the work of Korhonen (2002a, 2002b) is to develop a semantically driven approach to subcategorization frame hypothesis selection that can be used to improve large-scale subcategorization frame acquisition. The main idea underlying the approach is to leverage the well-known mapping between syntax and semantics, inspired by Levin's (1993) work. Korhonen uses statistics over semantic classes of verbs to smooth a distribution of subcategorization frames and then applies a simple frequency cutoff to select the most reliable subcategorization frames. Her work is related to ours in several ways. First, the automatic acquisition task leverages correspondences between syntax and semantics, particularly clear in the organization of the verb lexicon, similarly to Merlo and Stevenson (2001). Some of our current results are also based on this correspondence, as we assume that the notion of argument is a notion at the interface of the syntactic and semantic levels, and participates in both, determining not only the valency of a verb but also its subcategorization frame. Our work confirms the results reported in Korhonen (2002a), which indicate that using word classes improves the extraction of subcategorization frames. Differently from Korhonen, however, we do not allow feedback between levels. In her work, syntactic similarity of verbs' subcategorization sets based on an external resource (LDOCE codes) are used to determine a semantic classification of verbs—or rather to partially reorganize Levin's classification. This semantic classification is then used to collect statistics that are used to smooth a subcategorization distribution. In our work, instead, we do use WordNet in some places to give us information on verb classes, but we never use explicit semantic information on the set of verb subcategorization frames to determine the notion of argument or adjunct.

Aldezabal et al. (2002) is another piece of work related to our current proposal. In this article, the distinction between arguments and adjuncts is made to determine the subcategorization frames of verbs for a language, Basque, for which not many developed resources exist. This illustrates another use of the distinction between argument and adjuncts, which is not apparent when working on English.

6.3 On Learning Abstract Notions Using Corpus-based Statistical Methods

Learning arguments and adjuncts is an example of learning simplified semantic information by using syntactic and lexical semantic correlates. We learn the target concepts of arguments and adjuncts by using corpus-based indicators of their properties. It remains to be determined if we just learn correlates of a unified notion, or if the distinction between arguments and adjuncts is a clustering of possibly unrelated properties.

As explained in the introduction, native speakers' judgments on the argument and adjunct status of PPs are very unstable. No explanation is usually proposed of the fact that the tests of argumenthood are often difficult to judge or even contradict each other. As a possible explanation for the difficulty in pinpointing exactly the properties of arguments and adjuncts, Manning (2003) suggests that the notion of argument or adjunct is not categorical. The different properties of argument and adjuncts are not the reflex of a single grammatical underlying notion, but they can be ascribed to different mechanisms. What appears as a not entirely unified behavior is in fact better explained as separate properties.

The current article provides a representation that can support both the categorical and the gradient approach to the distinction between arguments and adjuncts. We have decomposed the notion of argument into a vector of features. The notion of argumenthood is no longer necessarily binary, but it allows several dimensions of variation, each potentially related to a different principle of grammar. In the current article, we have adopted a supervised approach to the learning task and adopted a binary classification. To pursue a line of reasoning where a gradient representation of the notion of argument is preferred, we would no longer be interested in classifying the vectorial information according to a predetermined binary or four-way target value, as was done in the supervised learning experiments. The appropriate framework would then be unsupervised learning, where several algorithms are available to explore the hidden regularities of a vectorial representation of clusters of PPs.

Whether a linguistic notion should be considered categorical or gradient is both a matter of empirical fact and of the explanatory power of the theory into which the notion is embedded. Assessing the strengths and weaknesses of these two approaches is beyond the scope of the current article.

7. Conclusions

We have proposed an augmentation of the problem of PP attachment as a four-way disambiguation problem, arguing that what is needed in interpreting prepositional phrases is knowledge about both the structural attachment site (the traditional noun–verb attachment distinction) and the nature of the attachment (the distinction of arguments from adjuncts). Practically, we have proposed a method to learn arguments and adjuncts based on a definition of arguments as a vector of features. Each feature is either a lexical element or its semantic class or it is a numerical representation of a diagnostic that is used by linguists to determine if a PP is an argument or not. We have shown in particular that using lexical classes as features yields good results, and that diagnostics based on linguistic theory improve the performance even further. We have also argued that the notion of argument does not help much in disambiguating the attachment site of PPs, indicating that the two notions are not closely correlated and must be established independently. We have performed a series of four-way classification experiments, where we classify PPs as arguments or adjuncts of a noun, and as arguments or adjuncts of a verb. We show that the classification performances are reasonably good for verb adjuncts, noun arguments, and noun adjuncts, independent of the learning algorithm. Classification performances of prepositional phrases that are arguments of the verb are poor if decision trees are used, but are greatly improved by the use of a large margin classifier. The features that appear to be most effective are lexical classes, thus confirming current linguistic theories that postulate that a verb's argument structure depends on a verb's lexical semantics (Levin 1993). Future work lies in further investigating the difference between arguments and adjuncts to achieve even finer-grained classifications and to model more precisely the semantic core of a sentence.

1. Appendix: PP Configurations

	Sequence of single PP attached to a verb	
Configuration	Structure	Example
Transitive	[vp V NP PP]	join board as director
Passive	[vp NP PP]	tracked (yield) by report
Sentential Object	[vp V NP PP]	continued (to slide) amid signs
Intransitive	[vp V PP]	talking about years
	Sequence of single PP attached to a noun	
Noun phrase	[np NP PP]	form of asbestos
Transitive	[vp V [np NP PP]]	have information on users
Transitive with two PPs, one attached to verb, other to noun	[vp V [np NP PP] PP]	dumped sacks of material into bin
Noun phrase with two PPs attached low	[np NP [pp P [np NP PP]]]	exports at end of year
Transitive verb with two PPs attached low	[vp V [np NP [pp P [np NP PP]]]]	lead team of researchers from institute
Transitive with two PPs, one attached to verb, other to other PP	[vp V NP [pp P [np NP PP]]]	imposed ban on all of asbestos
Intransitive with two PPs, one attached to verb, other to noun	[vp V [pp P [np NP PP]]]	appear in journal of medicine
Phrasal object with two PPs	[vp V NP [pp P [np NP PP]]]	continued (to surge) on rumors of buying
Passive form with two PPs	[vp V NP [pp P [np NP PP]]]	approved (request) by houses of Congress
	Sequence of 2 PPs attached to a verb	
Intransitive	[vp V PP PP]	grew by billion during week
Passive	[vp V NP PP PP]	passed (bill) by Senate in forms
Transitive	[vp V NP PP PP]	giving 16 to graders at school
	Sequence of 2 PPs attached to a noun	
Noun	[np NP PP PP]	sales of buses in October
Transitive	[vp V [np NP PP PP]]	meet demands for products in Korea

Acknowledgments

Most of this research was conducted thanks to the generous support of the Swiss National Science Foundation, under grant 1114-065328.01, while the second author was a master's student at the University of Geneva. We thank Eric Joanis for his precious help in constituting the database of PP tuples. We also thank Jamie Henderson for providing native-speaker judgments.

References

Aldezabal, Izaskun, Maxux Aranzabe, Koldo Gojenola, Kepa Sarasola, and Aitziber Atutxa. 2002. Learning argument/adjunct distinction for Basque. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon on Unsupervised Lexical Acquisition*, pages 42–50, Philadelphia, PA.

Argaman, Vered and Neal Pearlmutter. 2002. Lexical semantics as a basis for argument structure frequency bias. In Paola Merlo and Suzanne Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. John Benjamins, Amsterdam/Philadelphia, pages 303–324.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the Thirty-sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics (ACL-COLING'98)*, pages 86–90, Montreal, Canada.

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II style, Penn Treebank Project. Technical report, University of Pennsylvania, Philadephia.

Buchholz, Sabine. 1999. Distinguishing complements from adjuncts using memory-based learning. ILK, Computational Linguistics, Tilburg University.

Chang, Chih-Chung and Chih-Jen Lin. 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Collins, Michael and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 27–38, Cambridge, MA.

CoNNL. 2004. *Eighth Conference on Computational Natural Language Learning* (CoNLL-2004). Boston, MA.

CoNLL. 2005. *Ninth Conference on Computational Natural Language Learning* (CoNLL-2005). Ann Arbor, MI.

Dorr, Bonnie. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):1–55.

Fabre, Cécile and Didier Bourigault. 2001. Linguistic clues for corpus-based acquisition of lexical dependencies. In *Proceedings of the Corpus Linguistics Conference*, pages 176–184, Lancaster, UK.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Grimshaw, Jane. 1990. *Argument Structure*. MIT Press.

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Jackendoff, Ray. 1977. *X' Syntax: A Study of Phrase Structure*. MIT Press, Cambridge, MA.

Korhonen, Anna. 2002a. Semantically motivated subcategorization acquisition. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon on Unsupervised Lexical Acquisition*, pages 51–58, Philadelphia, PA, July.

Korhonen, Anna. 2002b. *Subcategorisation Acquisition*. Ph.D. thesis, University of Cambridge.

Levin, Beth. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, IL.

Manning, Christopher. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay, and Stephanie Jannedy, editors, *Probabilistic Linguistics*. MIT Press, pages 289–314.

Marantz, Alex. 1984. *On the Nature of Grammatical Relations*. MIT Press, Cambridge, MA.

Marcus, M., G. Kim, A. Marcinkiewicz, R. Macintyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 114–119, Plainsboro, NJ.

Marcus, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Merlo, Paola. 2003. Generalised PP-attachment disambiguation using corpus-based linguistic diagnostics.

In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 251–258, Budapest, Hungary.

Merlo, Paola, Matt Crocker, and Cathy Berthouzoz. 1997. Attaching multiple prepositional phrases: Generalized backed-off estimation. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 145–154, Providence, RI.

Merlo, Paola and Matthias Leybold. 2001. Automatic distinction of arguments and modifiers: The case of prepositional phrases. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 121–128, Toulouse, France.

Merlo, Paola and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on Wordnet. Technical report, Cognitive Science Laboratory, Princeton University.

Nielsen, Rodney and Sameer Pradhan. 2004. Mixing weak learners in semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 80–87, Barcelona, Spain, July.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105.

Phillips, William and Ellen Riloff. 2002. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 125–132, Philadelphia, PA.

Pollard, Carl and Ivan Sag. 1987. *An Information-based Syntax and Semantics*, volume 13. CSLI Lecture Notes, Stanford University.

Quinlan, J. Ross. 1993. *C4.5 : Programs for Machine Learning*. Series in Machine Learning. Morgan Kaufmann, San Mateo, CA.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Ratnaparkhi, Adwait. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Providence, RI.

Ratnaparkhi, Adwait, Jeffrey Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 250–255, Plainsboro, NJ.

Riloff, Ellen and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 49–56, Montreal.

Schütze, Carson T. 1995. PP Attachment and Argumenthood. *MIT Working Papers in Linguistics*, 26:95–151.

SENSEVAL-3. 2004. *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (SENSEVAL-3). Barcelona, Spain.

Srinivas, Bangalore and Aravind K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Stede, Manfred. 1998. A generative perspective on verb alternations. *Computational Linguistics*, 24(3):401–430.

Stetina, Jiri and Makoto Nagao. 1997. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 66–80, Beijing/ Hong Kong.

Swier, Robert and Suzanne Stevenson. 2005. Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-05)*, pages 883–890, Vancouver, Canada.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.

Villavicencio, Aline. 2002. Learning to distinguish PP arguments from adjuncts. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 84–90, Taipei, Taiwan.

Xue, Nianwen. 2004. Handling dislocated and discontinuous constituents in Chinese semantic role labelling. In *Proceedings of the Fourth Workshop on Asian Language Resources (ALR04)*, pages 19–26, Hainan Island, China.

Xue, Nianwen and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 88–94, Barcelona, Spain.

Yeh, Alexander. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference in Computational Linguistics (COLING 2000)*, pages 947–953, Saarbruecken, Germany.

Zhao, Shaojun and Dekang Lin. 2004. A nearest-neighbor method for resolving PP-attachment ambiguities. In *The First International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 545–554, Hainan Island, China.