Automatic Distinction of Arguments and Modifiers: the Case of Prepositional Phrases

Paola Merlo
Linguistics Department
University of Geneva
2 rue de Candolle
1211 Geneva 4, Switzerland
merlo@lettres.unige.ch

Matthias Leybold
American Management Systems
Weltpoststrasse 20
3000 Bern 15,
Switzerland
Matthias.Leybold@ams.com

Abstract

The automatic distinction of arguments and modifiers is a necessary step for the automatic acquisition of subcategorisation frames and argument structure. In this work, we report on supervised learning experiments to learn this distinction for the difficult case of prepositional phrases attached to the verb. We develop statistical indicators of linguistic diagnostics for argumenthood, and we approximate them with counts extracted from an annotated corpus. We reach an accuracy of 86.5%,over a baseline of 74% , showing that this novel method is promising in solving this difficult problem.

Introduction

The ability to automatically distinguish between arguments and modifiers is necessary for the automatic acquisition of important lexical knowledge, such as subcategorisation frames and argument structures, which is used in parsing, generation, machine translation, information extraction (Srinivas and Joshi, 1999; Stede, 1998; Dorr, 1997; Riloff and Schmelzenbach, 1998). Yet, few attempts have been made to perform this distinction automatically. In this paper, we present results related to what has been found to be a particularly difficult instance of this problem: the case of prepositional phrases (PPs) attached to the verb. Previous work shows that this distinction is more difficult for PPs than for other parts of speech (Buchholz, 1999), and also that attachment to a verb in general is less accurately performed than attachment to a noun (Hindle and Rooth, 1993).

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 modifiers 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, while a modifier 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, while it is a modifier if its interpretation remains relatively constant when associating with different heads, (Grimshaw, 1990, 108). These semantic differences give rise, among others, to some observable distributional consequences: for a given interpretation, a modifier can co-occur with a relatively broad range of heads, while arguments are limited to co-occurrence with a (semantically restricted) class of heads (Pollard and Sag, 1987, 136).

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

a) Maria is a student **of physics**

b) Maria is a student from Phoenix

In example a), 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 b) instead, the PP predicates a different property of the student, namely her geographical origin, which is not implied by the head *student*.

a) Kim camps/jogs/meditates on Sunday.

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

In example a) 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 b). 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 surface in observable syntactic differences which can be used to elicit acceptability judgments. However, the linguistic diagnostics that are used to determine whether a PP is a modifier 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-modifier distinction, which we approximate statistically and estimate using corpus counts. The diagnostics are then automatically combined in a decision tree induction algorithm. A detailed analysis of the behaviour of the classifier suggests that the diagnostic features do capture defining properties of arguments and modifiers.

The Linguistic 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 copular paraphrases, pro-form replacement and wh-extraction. We choose four diagnostics that tap more directly into the semantic properties of the sentence and can be captured by simple statistical concepts, easily estimated in a corpus. The diagnostics are head dependence, optionality, iterativity and ordering.

Head Dependence Arguments depend on their lexical heads, because they form an integral part of the phrase. Modifiers do not. Consequently, PP-arguments can only appear with the specific verbal head by which they are lexically selected, while PP-modifiers can co-occur with a far greater range of different heads than arguments, as illustrated in the example sentences above. We capture this insight by counting the number of different verbs that co-occur with a given PP in a corpus, as indicated in (1). A low number indicates argument status, while a high number indicates modifier status.

$$opt_{PP} = |\{v_1, v_2, v_3, \ldots, v_n\}_{PP}| \qquad (1)$$

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

a) John put the book **in the room**.
b) *John put the book.

c) John saw/read the book in the room
d) John saw/read the book.

Thus we expect that the predictive power of a verb about its complements will be greater for arguments than for modifiers. This insight can be captured by the conditional probability of a PP given a particular verbal head, as indicated in (2). We expect to find larger values for arguments than for modifiers.

$$P(p, n_2|v) = \frac{C(v, p, n_2)}{C(v)} \qquad (2)$$

Notice that this diagnostics can only be interpreted as a statistical tendency, and not as a strict test, because not all arguments are obligatory (but all modifiers 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 PPs, instrumentals have been argued to be arguments (Schütze, 1995). While keeping these exceptions in mind, we maintain optionality as a valid diagnostic here.

Iterativity and Ordering Arguments cannot be iterated and must be adjacent to the selecting lexical head, as illustrated below (arguments in bold):

a) *Chris rented the gazebo **to yuppies, to libertarians.**
b) Kim met Sandy in Baltimore in the hotel lobby in a corner.

Consequently, in a sequence of several PPs only the first one can be an argument, while the others must be modifiers. The correlate of this test of iterativity and ordering, then, is simply whether a given PP is found in second (or higher) position in a corpus of multiple PP sequences. This would indicate a modifier status.

In conclusion, the diagnostics of head dependence, optionality, iterativity and ordering are promising indicators of argumenthood as they make clearly different predictions for arguments and for modifiers, and they are approximated by an indicators which can be estimated in a sufficiently large corpus.

Materials and Method

Corpora Our corpora correspond to the subsets of verb attachments extracted from two publicly available PP-attachment corpora (Merlo et al., 1997) extracted from the Penn Treebank (Marcus et al., 1993). One corpus contains data encoding information for attachment of single PPs in the form of four head words (verb, object noun, preposition and noun inside the preposition) for each instance of PP attachments found in the corpus.[1] We also use an auxiliary corpus of 264 sequences of three PPs, where each data item consists of the two PPs following the head noun and verb and of the third preposition. A portion of the first corpus – the single PP corpus – was kept aside for testing, and was never used to develop the counts described below.

Counts Head dependence is approximated in two ways. First, by the size of the set of verbs that can co-occur with a given PP, see (1) above. Clearly, this measure will depend on finding exactly the same PP-internal noun to match, and will suffer from sparse data. We implement then a second variant where we cluster PPs according to the semantic content of the PP-internal nouns. For comparison, the semantic labelling has been done manually, and automatically, using Wordnet 1.6.

In the manual labelling, the PP-internal nouns are labelled semantically, with two sets of features. The first feature is meant to be a coarse semantic feature that reflects the syntactic behaviour of the labelled noun. Its values are : *activity, animal, attribute, concept, discipline, environment, event, human, institution, name, object, place, quantity, state, state-of-mind, substance, time.* The second set of features is meant to be more finely grained, and it has values such as *period, punctual, agent, private, public, business, institution, language, nation, object, person, amount, unit.*

To compare the effectiveness of this labelling and to attempt a more scalable approach, we developed an automatic method based on on the lexicographer's classes in Wordnet 1.6 (Miller et al., 1990). Nouns are classified in 25 different classes, among which, for example, *animal, artifact, attribute, body, cognition, communication, event, feeling, food, location, motive, person, plant, process, quantity, relation, shape, substance.* This second classification required selecting one particular Wordnet sense for those polysemous nouns being classified. To perform this selection automatically in a simple way, we always chose the first sense in Wordnet, which is the most frequent.

Optionality is measured in two variants. First, it is calculated as a conditional probability based on simple word counts in the corpus of single PPs, as indicated in (2) above. Second, we also implement a variant that relies on verb classes instead of individual verbs (see 3), to address the problem of sparse data. Verbs were grouped into classes using Levin's classification (Levin, 1993) and disambiguated by hand, using intuition.

$$P(p, n_2|vcl) = \frac{C(vcl, p, n_2)}{C(vcl)} \qquad (3)$$

For the iterativity measure, counts were collected indicating whether a given PP in first position in the single PP corpus had been found in second position in the multiple PP corpus, an

[1] This corpus differs from the one described in (Ratnaparkhi et al., 1994). In the current corpus, all instances of PPs immediately followed the verb and noun in the Penn Treebank, while in Ratnaparkhi et al.'s this is not always the case. As shown in (Merlo et al., 1997), the distance of the PP from the head verb and noun affects attachment preferences, and therefore PPs in later positions should not be considered good examples of PP attachment in first position.

indication of modifier status. We avoided the problem of sparse data using a backed-off algorithm (Collins and Brooks, 1995). For more detail on all these counts, see (Leybold, 2001).

The Experiment

The Input Data Each input vector contains 17 training features – the four lexical heads, and all the different variants of the implementation of the diagnostics – and one goal feature, indicating the argument or modifier status. We illustrate these in turn.

– *v, n1, p,, n2* indicate the four lexical heads (verb, object, preposition, and PP-internal noun);

– *vcl* indicates the verb class to which the verb in the tuple to be classified belongs, following (Levin, 1993), as discussed above for the optionality feature;

– *f1* and *f2* are the two types of hand-developed features used to cluster the PP-internal nouns semantically, as discussed above for the head dependence feature;

– *hdep1, hdep2,* and *hdep3* are the measures of head dependence using the features *f1* and *f2* to cluster the PPs, and using verb tokens, verb types and verb classes, respectively, to calculate the measure;

– *wncl* indicates the features extracted from Wordnet to cluster the nouns inside the PPs;

– *hdepwn1, hdepwn2,* and *hdepwn3* are the measures of head dependence using the *wncl* to cluster the PPs, and using verb tokens, verb types and verb classes, respectively, to calculate the measure;

– *opt1* is the measure of optionality using the individual verbs, $P(p, n_2|v)$;

– *opt2* is the measure of optionality using verb classes, $P(p, n_2|vcl)$;

– *iter* is the measure of iterativity;

– *status* is the goal predicate for the task, which can have the value 0 (arg) or 1 (mod).

Notice that the Penn TreeBank does not distinguish explicitly between arguments and modifiers, but rather it uses a combination of structural information and functional tags, indicating for example that the PP is a manner PP or a locative PP. We mapped those PPs that have no functional tags to arguments, and those that have functional tags, except CLR, to modifier. The tag CLR stands for closely related, and it would therefore seem to indicate that the PP should be considered an argument (Buchholz, 1999), To confirm the mapping of CLRs to arguments, we performed a distributional analysis of CLR similarity to the other two tags. Specifically, the distributional behaviour of the CLR-tagged elements was compared to that of arguments and modifiers. For the three groups (arguments, CLRs and modifiers) a "signature" based on conditional probability was developed. The average conditional probability of the first noun given the verb and of the preposition given the verb, and the average conditional probability of the second noun given the preceding preposition and the verb were calculated. Results confirm that CLRs are more similar to arguments than to modifiers.

Method We tested the results in two different ways. First we used a single training corpus containing 3692 exemplars, while 400 items were used for testing. We use the C5.0 Decision Tree Induction Algorithm (Quinlan, 1992). The baseline accuracy of this task is of 74% calculated by performing the classification using only the feature *preposition*. The default assignment is modifier attachment which reaches 52% accuracy. Second, we verified the generality of the results in a difference way by performing a cross-validation. This was done by partitioning the training examples in 10 subsets and using 9 subsets to train, rotating the composition of each of the 10 training sets, yielding 3321 training examples. The testing is always done on the same 400 examples. Notice that this procedure is slightly different from the usual way of performing cross-validation, where all the data are used. In our case, it is necessary to always use a separate testing set because the 3692 examples in the training set were directly inspected to generate the values for some of the features, such as *f1* and *f2*. The testing set instead consists of examples that were never used during development.

Results Table 1 indicates that the head dependence feature in all its variants is quite effective and gives much better performance than the other features. On the other hand, the iterativity features is not useful, perhaps because it was calculated using too small an auxiliary corpus of multiple PPs. The optionality feature is better than the default assignment, but not very

Feature Used	Accuracy
hdep3	67.5
hdep2	67.2
hdep1	66.8
hdepwn1	65.5
hdepwn2	65.5
hdepwn3	64.0
opt2	62.8
opt1	57.0
iter	52.0
default (mod)	52.0

Table 1: Results using variants of each diagnostic individually.

effective. We discuss possible interpretations of the results below.

We have run experiments with very many different feature combinations. A summary of the most interesting patterns of results are indicated in Table 2. The table indicates the feature used, the percentage of accuracy achieved using the entire training set (3692 examples), and the mean and standard error of the results achieved by using 10 different training sets. Significance of the difference in accuracy of two given sets of features can be calculated by taking each accuracy plus or minus the double of the standard error (at the $p<.05$ level). If the two ranges overlap, then the difference in accuracy is not significant. For example, the difference between the first and third line is not significant.

The best performance (line 1) is achieved by using a selection of the features, mixing some lexical information (the preposition), some general semantic information (the verb class and the PP-internal noun class derived from Wordnet) and only one of the variants of each of the diagnostic features about argumenthood (with minimal differences depending on which variant is chosen). The following lines (2–9) indicate the respective contribution of each of these features. By comparing the rows in the table to the top one, we can get information on how features perform, by looking at how performance varies when features are not used for the classification. Line 2 shows that the feature iterativity is in fact not useful at all, as already indicated by the individual feature results. Differently from the individual feature results, head dependence is not as useful as optionality in combination (line 3 and 5). Class information of both noun and verb, the latter especially, are quite effective (lines 4 and 6). Line 7 indicates the accuracy that can be achieved simply using the features that are already available in the current corpus of quadruples developed for PP attachment. Comparing the results using semantic classes and the argumenthood features (line 1-6) to line 7, we see that the contribution of the argumenthood features is positive, and on a par with the semantic information provided by grouping nouns into classes, probably because all these features are indicators of the same underlying semantic notions. Nonetheless, it is clear that the lexical features, especially those related to verb and preposition, still give the biggest reduction in error rate (lines 6, 8 and 9).

Separate calculation of precision and recall for arguments and modifiers indicate that arguments are individuated more precisely, while modifiers have better recall (Arguments: prec. 88.8%, prec. 82.3%. Modifiers: 84.7%, prec. 90.4%). This indicates a tendency to overassign the value of modifier, the default value. A separate analysis for each of the most frequent prepositions gives the following accuracies: *in* = 89.3%, *as* = 93.0%, *with* = 86.5%, *from* = 70.5%, *into* = 84.2%, *on* = 77.2%, *at* = 90.2%, *for* = 80.3%, *to* = 84.6%. Accuracies for prepositions that prefer an argument attachment are comparable to those of prepositions that prefer a modifier attachment of the PP.

Discussion

The current work explores a method to classify prepositional phrases as arguments or modifiers. It compares favourably to the only other study of which we are aware on this topic (Buchholz, 1999). The two studies are however not directly comparable, as Buchholz attempts a much larger-scale classification of all types of constituents. Buchholz reports an accuracy of 77% for PPs, but she does not restrict the attachment site to verbs.

The results reported above suggest that, like for part-of-speech tagging, simple frequencies of lexical items go a long way to solving the problem, while more sophisticated attempts improve the accuracy of a few percentage points.

Features Used	Accuracy %	Xval Mean %	SE %
1. p, vcl, wncl, opt2, hdep3, iter	86.5	85.9	0.2
2. p, vcl, wncl, opt2, hdep3	86.5	85.9	0.2
3. p, vcl, wncl, opt2, iter	85.8	85.2	0.2
4. p, vcl, opt2, hdep3, iter	84.8	85.1	0.3
5. p, vcl, wncl, hdep3, iter	84.5	84.2	0.1
6. p, wncl, opt2, hdep3, iter	82.0	81.9	0.1
7. v, n1, p, n2	82.0	79.5	1.0
8. p, vcl	81.5	81.0	0.3
9. vcl, wncl, opt2, hdep3, iter	75.0	73.5	0.4

Table 2: Results using combinations of features.

The significance of this work then resides in the exploration of the idea that the distinction between arguments and modifiers can be approached by capturing linguistic concepts statistically, and in the investigation of the relevant linguistic concepts. More specifically this work suggests two conclusions: First, while co-occurrence between individual words is very effective, sometimes co-occurrence between word classes is even more effective. Word classes capture syntactic and semantic regularities which help in disambiguation, while also providing an effective method to avoid sparse data. Second, this work supports the idea that complex distinctions, such as argument and modifier, are properly captured by a combination of linguistic diagnostics that measure both syntagmatic word co-occurrence – co-occurrence in the string –, such as optionality, and also dispersion across the set of words that can occur in a given position – paradigmatic co-occurrence –, such as head dependence.

In discussing the fact that they obtain worse results for modifiers than for arguments, (Hindle and Rooth, 1993) already mention the possibility that their lexical association measure is not appropriate for modifiers. This conclusion is supported by the observation that one of the crucial properties of modifiers is indeed that they are *not* selected by a lexical head. A measure that captures this property is unlikely to be based on word co-occurrence. [2] Thus, we propose that word co-occurrence measures capture properties of arguments, such as our optionality, while the defining property of modifiers are captured by paradigmatic associations, such as head dependence.

To verify that the diagnostic features do indeed capture these two dimensions of association between words, we performed some experiments, on a reduced set of features to see clearly the interaction of the features of interest. The logic is as follows. The measure of optionality is a measure of association of the verb and the PP, indicating how strongly the verb selects the PP. This measure will only show a statistical tendency. There could be cases of very high co-occurrences of certain verbal heads with certain PP modifiers, that are not due to the linguistic content of the sentence, but to non-linguistic, circumstantial influences. For example, take the sentence *The stock market rose 5% on Monday*. Sentences similar to this are very frequent in the Wall Street Journal. In this case, the optionality feature incorrectly classifies the PP as an argument. Thus, if the optionality measure really captures properties of arguments, we should find that for high values of the feature – indicating argument status – modifiers are misclassified as arguments. For the low values of the feature – which indicate modifier status – on the other hand, one should find the opposite tendency to misclassify arguments as modifiers. This difference in the classification due to optionality interacts with the head dependence feature. Consider again the example sentence above. In this case, the head dependence feature will correctly identify the PP as a modifier by the previous cri-

[2] We did in fact experiment with a measure of head dependence that calculated the conditional probability of a verb given a PP. This measure still captured the idea that in modifiers it is the PP that picks the verb, but it did not capture the idea of dispersion. This measure was not used as it does not perform as well as the other implementations of head dependence.

Features Used	High				Low		
		Arg	Mod			Arg	Mod
prep opt2 hdep3	Arg	26	9		Arg	23	7
	Mod	3	18		Mod	6	30
prep opt2	Arg	31	4		Arg	20	10
	Mod	6	15		Mod	6	30

Table 3: Contribution to classification of the head dependence feature when interacting with low and high optionality values, using three features.

Features Used	High				Low		
		Arg	Mod			Arg	Mod
opt2 hdep3	Arg	26	9		Arg	22	8
	Mod	5	16		Mod	16	20
opt2	Arg	35	0		Arg	17	13
	Mod	21	0		Mod	3	33

Table 4: Contribution to classification of the head dependence feature when interacting with low and high optionality values, using two features.

terion, as *on Monday* can co-occur with many different classes of verbs. In the case of arguments incorrectly misclassified as modifiers, again head dependence will interact, separating the actual arguments (low value for head dependence) from the errors (high value for head dependence).

To verify these hypotheses, we created two new test sets, which differ in the range of values of the feature *opt2*. One test set contains only high values (greater than 0.5) and one contains only low values (lower than 0.2). We then performed two runs of experiments. First, we constructed the classifier using three features: *prep, hdep3, opt2*, and then using only *prep, opt2.* We tested both these classifiers with the high value and low value optionality test sets. The difference in the classification between the full set of feature and the set of features without head dependence will illustrate the contribution of the head dependence feature. We then repeated the experiments using only the *hdep3, opt2* features, to create an even clearer experimental setting. The results of all these experiments are shown in the confusion matrices reported in Tables 3 and 4. In the two tables, the values assigned by the classifier are indicated vertically, while the actual values are indicated horizontally.

Compare the first to the second panel in the High column in Table 3. These are examples where the optionality feature has a high value, and therefore should show an overestimation of arguments, as they do (see the lower panel, where the feature *hdep3* is not used to classify). Adding the head dependence feature increases the assignment to modifiers in general, reducing the false positives for argument assignment, as expected. In the panels describing the examples with low optionality values, one can observe the opposite pattern. Here, it is the assignment to modifiers that is overestimated, in the case where head dependence is not used (40 assignments to modifier in total). This result is expected, because these are examples that have a low optionality value. Adding the feature indicating head dependence decreases the number of false positives for modifier assignment – the number of actual arguments that were incorrectly assigned modifier status, which go from 10 to 7 – but does not change the other results. Table 4 is even clearer. For high optionality values, the optionality feature alone assigns all examples to the argument status. Adding head dependence moves many of the argument assignments to modifier status. For the low optionality values, the optionality feature alone overestimates the modifier assignment and the addition of the head dependence feature moves many of the examples towards an argument assignment. Thus, we can conclude that the two features of optionality and head dependence show the pattern of errors which we would expect, if optionality were particularly sensitive to factors related to arguments, and head dependence were sensitive to modifiers.

Conclusion and Future Work

The present paper has shown that it is indeed possible to make progress towards automatically determining the argumenthood status of a verb-attached prepositional phrase. The maximum result of 86.5% over a baseline of 74% suggests that this investigation represents a successful attempt at translating linguistic knowledge into statistical measures. In particular, it shows that the best results are achieved by augmenting lexical and word class information with linguistic diagnostics that measure both syntagmatic word co-occurrence, such as optionality, but also paradigmatic dispersion, such as head dependence.

Several areas for improvement and extensions remain. One observation that can be drawn from these experiments is that argumenthood features are useful, but not as useful as the lexical features. A possible explanation for this fact, which also constitute an indication for future research, is that the feature are calculated very naively. It is possible that with more sophisticated measure one would achieve better performance. In particular, the measure of iterativity could be improved by collecting a larger data set of multiple PPs. The measure of head dependence could be improved by using a more complex measure of dispersion, based on entropy. Since all measures seem to greatly benefit from grouping the nouns and verbs into classes, the classification would likely be more accurate if more sophisticated ways of creating these classes were used.

References

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

Michael Collins 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.

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

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

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

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

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

Matthias Leybold. 2001. Automatic distinction of pp-arguments and pp-modifiers based on statistic implementations of linguistic criteria: a contribution to the problem of pp-attachment disambiguation. Master's thesis, University of Geneva.

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

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

Paola Merlo, 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.

George Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on Wordnet. Technical report, Cognitive Science Lab, Princeton University.

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

J. Ross Quinlan. 1992. *C4.5 : Programs for Machine Learning.* Series in Machine Learning. Morgan Kaufmann, San Mateo, CA.

A. Ratnaparkhi, J. Reynar, and S Roukos. 1994. A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 250–255.

Ellen Riloff 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.

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

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

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