Michael Niemann (1998) **Determining PP Attachment through Semantic Associations and Preferences.** In D. Estival (ed.) *Abstracts for the ANLP Post Graduate Workshop,* pp 25-32.

Koolp and Collins & B

they can be computatio

and determine which

While lexical preferen

(2) a. I [ponght [s

b. I [ponght

formula to score the po

semantic role. Collins

easily affect the sense

As shown by examples

This is a vital compone

Table 5: The five most frequent attachment sites for the primary object WordNet groups
(N= noun group, V= verb group)

OBJECT		1		2		3		4		5	
	Total	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**
PROPER NOUN	133	24	action N	11	person N	9	group N	8	creation V	5	motion V stative V communication V communication N
ARTIFACT	113	14	motion V	11	artifact N	10	action N	9	contact V	6	communication V
ACTION	109	12	action N	9	change V	8	stative V	7	communication V	5	social V possession V communication N
TIME	84	9	action N	8	stative V	6	time N	5	motion V change V possession V social V	4	body V communication V
COGNITION	75	8	change V	6	stative V action N	5	cognition N				
PERSON	61	6	stative V action N	5	communication V communication N	4	motion V person N	3	cognition N	*same no. for more than four groups*	

Table 6: Top fivee attachment sites for a selection of prepositions
(N= noun group, V= verb group)

PREP		1		2		3		4		5	
	Total	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**
in	250	29	action N	21	change V	18	stative V	15	person N	9	event N group N
to	167	23	action N	18	motion V	10	change V stative V	9	communication N	7	communication V
for	100	9	cognition N	8	communication N	7	possession V action N	6	stative V	4	social V
with	97	10	change V	9	communication V	8	action N	7	stative V	5	creation V artifact N
on	86	7	change V	6	creation V motion V action N	5	communication V artifact N phenomenon N	4	cognition V cognition N communication N	3	possession N attribute N

Table 7: Top fivee object nouns for a selection of prepositions

PREP		1		2		3		4		5	
	Total	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**	**No.**	**Group**
in	250	41	proper noun	27	artifact action time	16	cognition	14	location state	13	attribute
to	167	25	proper noun	16	artifact	15	cognition	13	location	12	action
for	100	18	action	11	artifact	10	person time	7	cognition possession	6	group
with	97	15	person	13	cognition	9	artifact	8	attribute	7	action communication
on	86	10	cognition action	9	artifact	8	communication	7	time	6	proper noun

also a preference for attachment to prepositions, *action* nouns and *motion* verbs for this semantic role. Therefore, the attachment preferences may certainly be due to a preference for a particular role.

(7) a. over to Phil (preposition/proper noun)
b. go to crossroads (*motion* verb/*artifact* noun)
c. circle to lawn (*motion* verb/ *location* noun)

However, a semantic role cannot be as easily defined for *with* PPs as these PPs have no clear role preferences. Except for the *artifact/artifact* pair (see (8)), it is hard to associate particular attachment-site/object pairs with specific roles. However, such a relationship may occur if there were more data to inspect.

(8) a. building with dome
b. room with dais

The preferences of the prepositions and the different relationships between the various object and attachment site groups, should be taken into consideration by parser. Therefore, it is expected that the most accurate parser will consider all the semantic and syntactic information about the PPs and their possible attachment sites.

4. Experiments wwith Parsing

The PP parser Lexass was developed to show that the accuracy of parsing can be improved through the use of the WordNet semantic associations and the data extracted from the corpus. For 372 of the 1097 PPs in the data, a complete list of possible attachment sites was added to the data by hand. Moving left through the sentences, all nouns were included as possible sites until a verb or the start of the sentence was encountered. For instance, while *establishment* is the correct attachment site for (6), *authorize*, a *communication* verb, is a possible attachment site. The job of Lexass is to use the 1097 example PP attachments from the corpus[3] and the WordNet groupings to determine which of these sites *in the future* should be attached to.

Table 8: A Selecttion of the Levels used byLexass

LEVEL 33	all data
LEVEL 38	ignore attachment site WordNet group
LEVEL 40	ignore object WordNet group
LEVEL 16	ignore preposition
default	select left-most attachment site

[3] Of course, not including the PP being attached.

For each PP, Lexass moves through a sequence of scoring levels, trying to select the most likely attachment site. Each level accesses a file that contains the OCCURRENCENUMS for various patterns (see Table 8). An OCCURRENCENUM corresponds to the number of occurrences in the data of certain attachment patterns.

For instance, LEVEL 33 examines the number of times in the data that a particular preposition with a particular object is attached to each possible attachment site. The site with the highest OCCURRENCENUM for this pattern is the site to which the PP is attached more frequently in the data than any other possible sites.

Other levels generalize the counting of the OCCURRENCENUMS. LEVEL 40 ignores the WordNet group for the preposition's object, only including its SYNCODE. Therefore, for LEVEL 40, Lexass would use the OCCURRENCENUMS for an *in* PP with any noun object attached to a *communication* verb or an *action* noun.

To handle small frequencies and similar OCCURRENCENUMS, a cut-off score was used. If the top OCCURRENCENUM for a level was not more than the cut-off, or the difference between the two highest OCCURRENCENUMS was not greater than the cut-off, then the next level in the sequence was attempted (see Table 9).

Table 9: The algorithm for Lexass

- Read the level from the sequence parameter.
  - Score each attachment site according to the level.
  - If there is a valid top score ,
    select that attachment score.
  - Else, go to next level of scoring.
- If no scoring levvels give valid top scores,
  use the default mmethod of selection.

The final level in every sequence was the default level of selection. The accuracy of the default level was used as a benchmark to compare all other selection methods against. By default, PPs were attached to left-most site in the sentence. As this is normally the highest site in the sentence structure, the default method resembles the method of Minimal Attachment, except that it does not actually check for minimality.

This scoring method eliminates another of the problems with Hindle & Rooth and Collins & Brooks. Their methodologies are restricted to deciding between PP attachment to a verb or attachment a noun and their scoring formulae refer to all possible attachment sites in the one formula. They do not try and handle phrases

Table 10: Correct % at individual levels

	All PPs		DOUBLES		5PREPS		5DOUBLES	
Level Includes	No. Correct	% Correct	No. Correct	% Correct	No. Correct	% Correct	No. Correct	% Correct
Default (1)	*206*	*55 %*	*151*	*63 %*	*138*	*56 %*	*101*	*62 %*
All data (33)	16	76 %	8	80 %	13	72 %	7	78 %
No Object (40)	41	64 %	28	72 %	31	57 %	22	66 %
No Site (38)	49	64 %	33	72 %	35	57 %	25	66 %
No Prep (16)	131	48 %	89	52 %	80	43 %	53	47 %

Table 11: Correctt % of entire sequences
(Sequence = Leveel/s, Default)

	All PPs		DOUBLES		5PREPS		5DOUBLES	
Sequence Includes	No. Correct	% of 372	No. Correct	% of 239	No. Correct	% of 246	No. Correct	% of 162
1	*206*	*55 %*	*151*	*63 %*	*138*	*56 %*	*101*	*62 %*
33, 1	214	58 %	156	65 %	143	58 %	105	65 %
40, 1	213	57 %	156	65 %	140	57 %	104	64 %
38, 1	216	58 %	156	65 %	142	58 %	103	64 %
16, 1	184	49 %	130	54 %	114	46 %	82	51 %
33,38,40,1	218	59 %	157	66 %	142	58 %	103	64 %

like (9). These PPs can be handled by Lexass as, by using OCCURRENCENUMs, it scores each possible site individually. Therefore, it is not restricted to any particular number of possible attachments.[4]

(9) The man in the store with a white hat on his head

The experiments tried a range of sequences. The accuracy of chosen attachments varied depending on the levels in the sequences. As well as parsing all 372 of the PPs, various subgroups of the PPs were also parsed separately. DOUBLES is all the PPs with only two possible attachment sites. This allows some sort of comparison to be made with Hindle & Rooth and Collins & Brooks. 5PREPS is the selection of PPs which contain the five most common prepositions. 5DOUBLES is the selection of PPs from 5PREPS that have two possible attachment sites.

While the best sequence [33,38,40,1] selects the correct attachment only 59% of the time, this is clearly better than the 55% accuracy of the default method (see Table 11). If the accuracy of the selections made at individual levels is investigated (see Table 10), a number of further points can be made. Not surprisingly, the most accurate level is LEVEL 33 (76%) as it includes all of the semantic data about the PP and its possible attachment sites. However, only 21 of the 372 PPs were attached by LEVEL 33.[5] This is due to small size of the training data. Any further studies should include a larger set of data. The small amount of data also resulted in the best cut-off score being 0. Therefore, at LEV 33, any attachment site/preposition/object pattern in the training data provides a valid OCCURRENCENUM.

In contrast, LEVEL 40 and LEVEL 38 are able to attach more of the PPs due to the generalization of the patterns they were counting in the data. They are not as successful as LEVEL 33 (only 64% accurate) but are a clear improvement on the default level (55%).

The failure of LEVEL 16 indicates the importance of the preposition in the parsing process. The preferences of the preposition must be considered when parsing. The attachment should be guided by the choice of preposition involved as it is the sense of the preposition that will establish the semantic role of the PP. While this is still partially dependent on other components of the sentence, it is a vital factor that should not be ignored.

If the results for DOUBLES, 5PREPS and 5DOUBLES are investigated, the same sort of accuracies occur. Furthermore, the 80% accuracy of LEVEL 33 for DOUBLES is comfortably similar to the accuracy reached by Hindle & Rooth and Collinss & Brooks.

To test the efficiency of this parsing methodology with non-sense tagged texts, the 372 PPs were reprocessed in such a way that all the WordNet senses were presumed to be ambiguous. Hence, the lexical

[4] The 372 PPs had on average 2.5 possible attachment sites. Therefore, guesswwork would choose the correct attachment 40% of the time.

[5] I.e. the top OCCURRENCENUM for these PPs was a valid score.

items were placed in the WordNet group of the first sense encountered in WordNet. When parsed, the sequence [33,1] was only 57% accurate, with LEVEL 33 reducing from 76% accuracy to 72%. While these results are greater than the default level's benchmark of 55% accuracy, there is a clear reduction in accuracy when there is word sense ambiguity. However, as indicated in Section 3, this ambiguity does not always affect which WordNet groups some words are placed in. Hence the methodology recommended by this paper may be suitable even when the WordNet sense of words is unknown.

5. Conclusion

The results from this research demonstrate that the semantic associations given as hypernym and troponym trees in WordNet V5.1 can be used to categorize lexical items when parsing PP attachment. Clear preference patterns have been shown to exist, demonstrating relationships between prepositions, their objects and their attachment sites. The use of these preference patterns in the parsing was reasonably successful, provided the preposition was always included in the pattern.

The results also support the use of partial parsers. Partial parsing can be used to provide the basic phrasal structures in a sentence then semantic preferences, like those described in this paper, can be used to determine the most suitable phrasal attachments, given their context. Due to the varying levels of preferences, a system of weights, like those used for neural networks, may be required in order to balance the preferences in some way.

6. Acknowledgmennts

Thanks to Dominique Estival for her guidance. Thanks to the Department of Linguistics & Applied Linguistics, University of Melbourne, for purchasing and allowing me to use the DSO corpus and the Department of Computer Science for the computer facilities used for this research.

7. References

Abney, Steven. 1997. *The SCOL Manual: Version 0.1b*. http://www.sfs.nphil.uniituebingen.de/~abney/.

Abney, Steven. 1991. Parsing by Chunks. In Robert Berwick, Steven Abney & Carol Tenny (Eds.), *Principle-Based Parsing*. Dordrecht: Kluwer Academic Publishers. Also http://www.sfs.nphil.uni-tuebingen.de/~abney/.

Collins, Michael & Brooks, James. 1995. Prepositional Phrase Attachment through a Backed-Off Model. *Proceedings of the Association for Computational Linguistics Third Workshop on Very Large Corpora*. Also http://www.lanl.gov/archhive /cmp-lg/9506021 .

Fellbaum, Christine. 1993. English Verbs as a Semantic Net. In *Five Papers on WordNet* (CSL Report 43, Revised). Also http:// www.cogsci.princeeton.edu/wn/ .

Hindle, Donald & Rooth, Mats. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1), 103-120. Also http:// www2.ims.uni-stuttgart.de/~mats/ .

Miller, George A; Beckwith, Richard; Fellbaum, Christian; Gross, Derek & Miller, Katherine. 1993. Introduction to WordNet: An On-Line Lexical Database. In *Five Papers on WordNet* (CSL Report 43, Revised). Also http://www.cogsci.princeton.edu /~wn/ .

Miller, George A. & Fellbaum, Christian. 1991. Semantic networks of English. *Cognition*: *International Journal of Cognitive Science*, 41(1-3). Reprinted Beth Levin & Steven Pinker (Eds.), *Lexical & Conceptual Semantics*. Cambridge: Blackwell Publishhers. 197-229.

Niemann, Michael. 1997. *Determining Prepositional Phrase Attachment by Semantic Association and Preferences*. Honours thesis. Department of Linguistics & Applied Linguistics, the University of Melbourne, Melbouurne.

Smith, George W. 1991. *Computers and Human Language*. New York: Oxforrd University Press.

Taraban, Roman & McClelland, James L. 1988. Constituent Attachment and Thematic Role Assignment in Sentence Processing: Influences of Content-Based Expectations. *Journal of Memory and Language*. 27, 597-632.