Probabilistic Unification-Based Integration Of Syntactic and Semantic Preferences For Nominal Compounds

Dekai Wu*
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720 U.S.A.
dekai@ucbvax.berkeley.edu

Abstract

In this paper, we describe a probabilistic framework for unification-based grammars that facilitates integrating syntactic and semantic constraints and preferences. We share many of the concerns found in recent work on massively-parallel language interpretation models, although the proposal reflects our belief in the value of a higher-level account that is not stated in terms of distributed computation. We also feel that inadequate learning theories severely limit existing massively-parallel language interpretation models. A learning theory is not only interesting in its own right, but must underlie any quantitative account of language interpretation, because the complexity of interaction between constraints and preferences makes *ad hoc* trial-and-error strategies for picking numbers infeasible, particularly for semantics in realistically-sized domains.

Introduction

Massively-parallel models of language interpretation —including marker-passing models and neural networks of both the connectionist and PDP (parallel distributed processing) variety—have provoked some fundamental questions about the limits of symbolic, logic- or rule-based frameworks. Traditional frameworks have difficulty integrating preferences in the presence of complex dependency relationships. In analyzing ambiguous phrases, for example, semantic information should sometimes override syntactic preferences, and vice versa. Such interactions can take place at different levels within a phrase's constituent structure, even for a single analysis. Massively-parallel models excel at integrating different sources of preferences in a natural, intuitive fashion; for example, connectionist models simply translate dependency constraints into excitatory or inhibitory links in relaxation networks (Waltz & Pollack 1985). Furthermore, massively-parallel models have shown remarkable ability to compute complex semantic preferences.

We argue that it is possible and desirable to give a more meaningful account of preference integration at a higher level, without resort to distributed algorithms. One could say that we are interested in characterizing the nature of the preferences, rather than how they might be efficiently computed. We do not claim that all properties of massively-parallel models can or should be described at this level. However, few language interpretation models take advantage of those properties that can only be characterized at the distributed level.

We also propose a quantitative theory that assigns an interpretation to the numbers used in our model. A quantitative theory explains the numbers' significance by defining the procedure by which the model—in principle, at least—can learn the numbers. Much of the mystique of neural networks is due to their potential learning properties, but surprisingly, few PDP and no connectionist models of language interpretation that we know of specify quantitative theories, even though numbers must be used to run the models. Without a quantitative theoretical basis, it seems unlikely that the network structures will generalize much beyond the particular hand-coded examples, if for no other reason than the immense room for variation in constructing such networks.

Case Study: Nominal Compounds

Nominal compounds exemplify the sort of phenomena modeled by interacting preferences. Nouns themselves are often homonymous—is *dream state* a sleep condition or California?—necessitating lexical ambiguity resolution. Structural ambiguity resolution required for nested nominal compounds, which have more than one parse; consider [*baby pool*] *table* versus *baby* [*pool table*]. Lexicalized nominal compounds necessitate syntactic preferences, while semantic preferences are needed to guide semantic composition tasks like frame selection and case/role binding, as nominal compounds nearly always have multiple possible meanings. Traditionally, linguists have only classified nominal com-

*Many thanks to Robert Wilensky and Charles Fillmore for helpful discussions, and to Hans Karlgren and Nigel Ward for constructive suggestions on drafts. This research was sponsored in part by the Defense Advanced Research Projects Agency (DoD), monitored by the Space and Naval Warfare Systems Command under N00039-88-C-0292, the Office of Naval Research under contract N00014-89-J-3205, and the Sloan Foundation under grant 86-10-3.

PREFERRED PARSE	COMPETING LEXICALIZED COMPOUNDS		COMPETING LEXICALIZED AND IDENTIFICATIVE COMPOUNDS
	First compound more lexicalized	Second compound more lexicalized	
N N N	kiwi fruit juice LEXICALIZED LEXICALIZED	navel orange juice LEXICALIZED LEXICALIZED	afternoon rest area IDENTIFICATIVE LEXICALIZED
N N N	New York state park LEXICALIZED LEXICALIZED	baby pool table LEXICALIZED LEXICALIZED	gold watch chain LEXICALIZED IDENTIFICATIVE

Figure 1. Nominal compounds requiring integration of semantic preferences.

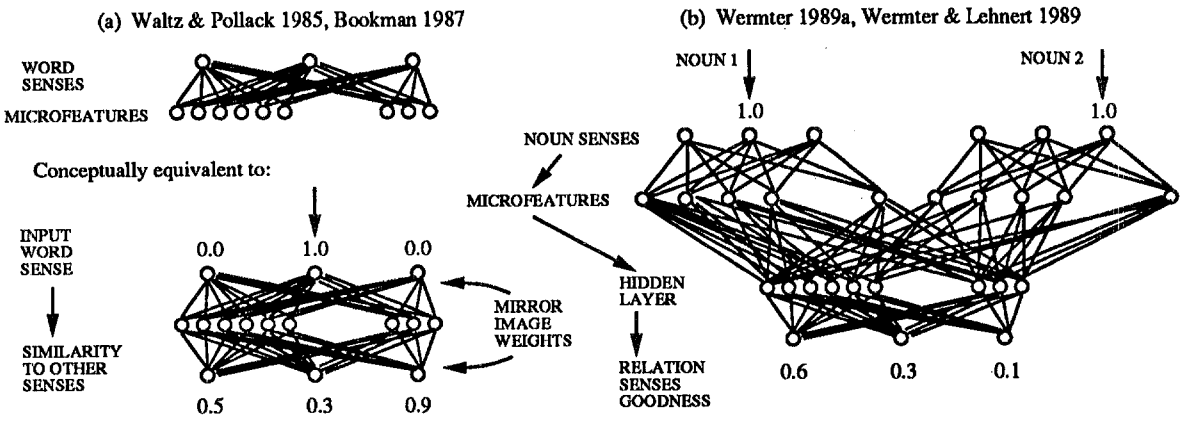

Figure 2. PDP semantic similarity evaluators.

pounds according to broad criteria such as part-whole or source-result relationships (Jespersen 1946; Quirk *et al.* 1985); several large-scale studies have provided somewhat finer-grained classifications on the order of a dozen classes (Lees 1963; Levi 1978; Warren 1978). However, the emphasis has been on predicting the possible meanings of a compound, rather than predicting its preferred meaning. An exception is Leonard's (1984) rule-based model which, however, only produces fairly coarse interpretations with medium (76%) accuracy.

We distinguish three major classes of nominal compounds: lexicalized (conventional), such as *clock radio*; identificative, such as *clock gears*; and creative, such as *clock table*. Both identificative and creative compounds are novel in Downing's (1977) sense; they differ in that an identificative compound serves to identify a known (but hitherto unnamed) semantic category, whereas to interpret a creative compound requires constructing a new semantic structure. There is a bias to use the most specific pre-existing categories that match the compound being analyzed, syntactic or semantic. Precedence is given to a conventional parse if one exists, then a parse with an identificative interpretation, and lastly a parse with a creative interpretation. However, this "Maximal Conventionality Principle" can easily be overruled by global considerations arising from the embedding phrase and context. Figure 1 shows examples where two conventional compounds compete, and where global considerations cause an identificative compound to be preferred over a competing conventional compound. These cases require integration of quantitative syntactic and semantic preferences, since non-quantitative integration schemes (e.g., Marcus 1980; Hirst 1987; Lytinen 1986) do not discriminate adequately between the alternative analyses.

What Do Massively-Parallel Models Really Say?

One use of massive parallelism is to evaluate the similarity or compatibility between two concepts in order to generate semantic preferences. Similarity evaluators usually employ PDP networks where semantic concepts are internally represented as distributed activation patterns over a set of "microfeatures". Conceptually, the network in Figure 2a gives a similarity metric between a given concept and every other concept, computed as the weighted sum of shared microfeatures.[1] Likewise, the hidden layer in Figure 2b computes the goodness of every possible relation between the given pair of nouns. In non-massively-parallel terms, what such nets do is capture statistical dependencies between concepts, down to the granularity of the chosen "microfeatures". A probabilistic feature-structure formalism employing the same granularity of features should be able to capture the same dependencies.

Connectionist models are often used to integrate syntactic and semantic preferences from different information sources (Cottrell 1984, 1985; Wermter 1989b; Wermter & Lehnert 1989). Nodes represent

[1] Ignoring Bookman's persistent activation, which simulates recency-based contextual priming.

hypotheses about word senses, parse structures, or role bindings; links represent either supportive or inhibitory dependencies between hypotheses. The links constrain the network so that activation propagation causes the net to relax into a state where the hypotheses are all consistent with one another. The most severe problem with these models is the arbitrariness of the numbers used; Cottrell, for example, admits "weight twiddling" and notes that lack of formal analysis hampers determination of parameters. In other words, although the networks settle into consistent states, there is no principle determining the probability of each state.

McClelland & Kawamoto's (1986) PDP model learns how syntactic (word-order) cues affect semantic frame/case selection, yielding more principled preference integration. Like the PDP similarity evaluators, however, the information encoded in the network and its weights is not easily comprehended.

Previous non-massively-parallel proposals for quantitative preference integration have used non-probabilistic evidence combination schemes. Schubert (1986) suggests summing "potentials" up the phrase-structure trees; these potentials derive from salience, logical-form typicality, and semantic typicality conditions. McDonald's (1982) noun compound interpretation model also sums different sources of syntactic, semantic, and contextual evidence. Though qualitatively appealing, additive calculi are liable to count the same evidence more than once, and use arbitrary evidence weighting schemes, making it impossible to construct a model that works for all cases. Hobbs *et al.* (1988) propose a theorem-proving model that integrates syntactic constraints with variable-cost abductive semantic and pragmatic assumptions. The danger of these non-probabilistic approaches, as with connectionist preference integrators, is that the use of poorly defined "magic numbers" makes large-scale generalization difficult.

A Probabilistic Unification-Based Preference Formulation

We are primarily concerned here with the following problem: given a nominal compound, determine the ranking of its possible interpretations from most to least likely. The problem can be formulated in terms of unification. Unification-based formalisms provide an elegant means of describing the information structures used to construct interpretations. Lexical and structural ambiguity resolution, as well as semantic composition, are readily characterized as choices between alternative sequences of unification operations.

A key feature of unification—especially important for preference integration—is its neutrality with respect to control, i.e., there is no inherent bias in the order of unifications, and thus, no bias as to which choices take precedence over others. Although nominal compound interpretation involves lexical and structural ambiguity resolution and semantic composition, it is not a good idea to centralize control around any single isolated task, because there is too much interaction. For example, the frame selection problem affects lexical ambiguity resolution (consider the special case where the frame selected is that signified by the lexical item). Likewise, frame selection and case/role binding are two aspects of the same semantic composition problem, and structural ambiguity resolution depends largely on preferences in semantic composition.

Thus we turn to unification for a clean formulation of the problem. Three classes of feature-structures are used: syntactic, semantic, and constructions. The *construction* is defined in Fillmore's (1988) Construction Grammar as "a pairing of a syntactic pattern with a meaning structure"; they are similar to signs in HPSG (Pollard & Sag 1987) and pattern-concept pairs (Wilensky & Arens 1980; Wilensky *et al.* 1988). Figure 3 shows a sample construction containing both syntactic and semantic feature-structures.[2] Typed feature-structures are used: the value of the special feature TYPE is a type in a multiple-inheritance type hierarchy, and two TYPE values unify only if they are not disjoint. This allows (1) easy transformation from semantic feature-structures to more convenient frame-based semantic network representations, and (2) efficient encoding of partially redundant lexical/syntactic categories using inheritance (see, for example, Pollard & Sag 1987; Jurafsky 1990). Our notation is chosen for generality; the exact encoding of signification relationships is inessential to our purpose here.

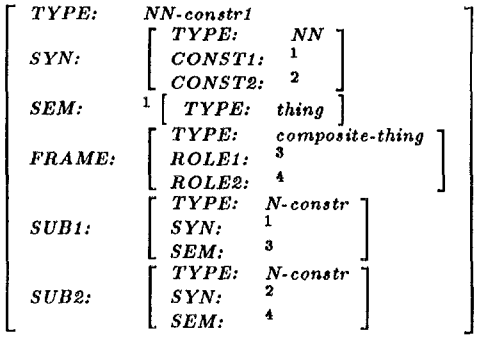

Figure 3. A nominal compound construction.

Given a nominal compound (of arbitary length), an *interpretation* is defined as an instantiated construction—including all the syntactic, semantic, and sub-construction f-structures—such that the syntactic structure parses the nominal compound, and the semantic structure is consistent with all the (sub-)constructions. Figure 4 shows an interpretation of *afternoon rest*. Given this framework, lexical ambiguity resolution is the selection of a particular sub-construction for a lexical item that matches more than one construction, structural ambiguity resolution is the selection between alternative syntactic f-structures, and semantic composition is the selection between alternative semantic f-structures. In each case we must be be able to compare alternative interpretations and determine the best.

Before discussing how to compare interpretations, let us briefly consider the sort of information available. We extend the unification paradigm with a function f that returns the relative frequency of any category in the type hierarchy, normalized so that for any category cat, $f(cat) = P[cat(x)]$ where x is a

[2]Ordering constraints are omitted in this paper.

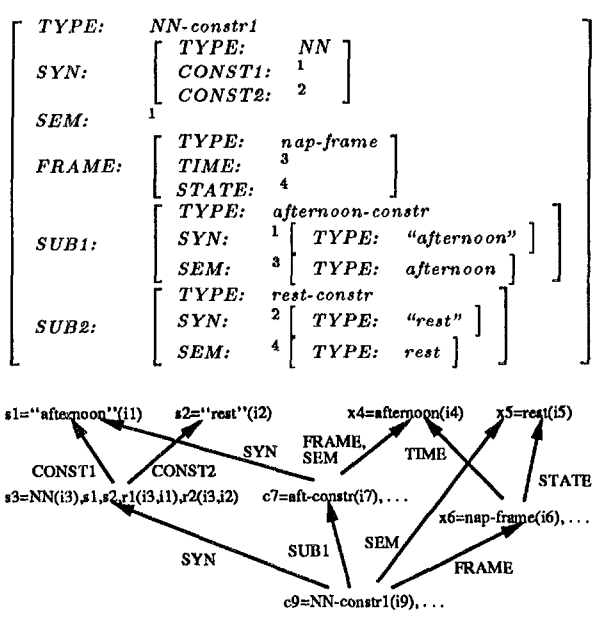

Figure 4. Bracket and graph representations of an interpretation of "afternoon rest".

random variable ranging over all categories. For semantic categories, this provides the means of encoding typicality information. For syntactic categories and constructions, this provides a means of encoding information about degrees of lexicalization. Since f is defined in terms of relative frequency, there is a learning procedure for f: given a training set of correct interpretations, we need only count the instances of each category (and then normalize).

The probabilistic goodness metric for an interpretation is defined as follows: the goodness of an interpretation is the probability of the entire construction given the input words of the nominal compound, e.g.,

$$P[+c_9| + s_1, +s_2] = P[\ \textit{NN-constr1}(i_9)|\ \textit{"afternoon"}(i_1) \wedge \textit{"rest"}(i_2)].$$

The metric is global, in that for any set of alternative interpretations, the most likely interpretation is that with the highest metric.

As a simplified example of computing the metric, suppose the feature graph of Figure 4 constituted a complete dependency graph containing all candidate hypotheses (actually an unrealistic assumption since this would preclude any alternative interpretations). For each pair of connected nodes, the conditional probability of the child, given the ancestor, is given by the ratio of their relative frequencies (Figure 5a). The metric only requires computing the probability of c_9 (Figure 5b).[3] Nodes are clustered into multi-valued compound variables as necessary to eliminate loops, to ensure counting any piece of evidence only once (Figure 5c).

The conditional probability vectors $P[+c_9|z_i]$ and $P[z_i| + s_1, +s_2]$ are computed using the *disjunctive interaction* model:[4]

$$P[+c_9| + s_3, +c_7, +c_8] = 1 - P[\neg c_9| + s_3] \cdot P[\neg c_9| + c_7] \cdot P[\neg c_9| + c_8]$$
$$P[+c_9| + s_3, +c_7, \neg c_8] = 1 - P[\neg c_9| + s_3] \cdot P[\neg c_9| + c_7]$$
$$P[+c_9| + s_3, \neg c_7, +c_8] = 1 - P[\neg c_9| + s_3] \cdot P[\neg c_9| + c_8]$$
$$P[+c_9| + s_3, \neg c_7, \neg c_8] = 1 - P[\neg c_9| + s_3]$$
$$P[+c_9|\neg s_3, +c_7, +c_8] = 1 - P[\neg c_9| + c_7] \cdot P[\neg c_9| + c_8]$$
$$P[+c_9|\neg s_3, +c_7, \neg c_8] = 1 - P[\neg c_9| + c_7]$$
$$P[+c_9|\neg s_3, \neg c_7, +c_8] = 1 - P[\neg c_9| + c_8]$$
$$P[+c_9|\neg s_3, \neg c_7, \neg c_8] = 1 - 1$$

$$P[+s_3, +c_7, +c_8| + s_1, +s_2] = P[+s_3| + s_1, +s_2] \cdot P[+c_7| + s_1, +s_2] \cdot P[+c_8| + s_1, +s_2] = \{1 - P[\neg s_3| + s_1] \cdot P[\neg s_3| + s_2]\} \cdot \ldots$$
$$P[+s_3, +c_7, \neg c_8| + s_1, +s_2] = \ldots$$

Finally, we compute $P[+c_9| + s_1, +s_2]$ by conditioning on the compound variable Z and taking the weighted average of $P[+c_9|Z, +s_1, +s_2]$ over all states of Z:

$$\sum_i P[+c_9|z_i, +s_1, +s_2]P[z_i| + s_1, +s_2] = \sum_i P[+c_9|z_i]P[z_i| + s_1, +s_2].$$

Both syntactic and semantic preferences are taken into account. The influence of semantic preferences is encoded in the conditional probabilities $P[+c_9| + c_7]$ and $P[+c_9| + c_8]$.[5] The loops in the original dependency graph correspond to support for the interpretation via both syntactic and semantic paths. A more complex example demonstrating structural ambiguity resolution is shown in Figure 6; here an afternoon rest schema produces a semantic preference that overrides a syntactic preference arising from weak lexicalization of the nominal compound *rest area*.[6]

A major unsolved problem with this approach is *specificity selection*. This is a well-known trade-off in classification models: the more general the interpretation, the higher its probability is; whereas the more specific the interpretation, the greater its utility and the more informative it is. The probabilistic goodness metric does not help when comparing two interpretations whose only difference is that one is more general than the other.[7] In our initial studies we attempted to handle this trade-off using thresholded marker-passing techniques (Wu 1987, 1989), but we are currently investigating a stronger *utility*

[3] A natural language processing system needs to propagate probabilities to the semantic hypotheses as well, in order to make use of the interpreted information.

[4] Justification for the disjunctive interaction model is beyond our scope here; it is a standard approximation used to complete the probability model in cases where is infeasible to gather or store full conditional probability matrices for all input combinations (see Pearl 1988). Heavily biased conditional probability matrices that cannot be satisfactorily approximated by disjunctive interaction can sometimes be handled by forming additional categories. The apparent schema-organization of human memory may well arise for the same reason.

[5] These conditional probabilities cannot be derived solely from frequency counts since c_9 is an instance of a novel category—the category of "afternoon rest" constructions denoting a nap—with zero previous frequency. Instead, the conditional probabilities $P[+c_9| + c_7]$ and $P[+c_9| + c_8]$ are a function of the ancestral conditional probabilities $P[+s_3|+s_1]$, $P[+s_3|+p_2]$, $P[+x_6|+x_4]$, and $P[+x_6|+x_5]$ plus the disjunctive interaction assumption.

[6] Note that (a) and (b) are two partitions of the same dependency graph.

[7] Norvig (1989) has also noted the competition between probability and utility in the context of language interpretation.

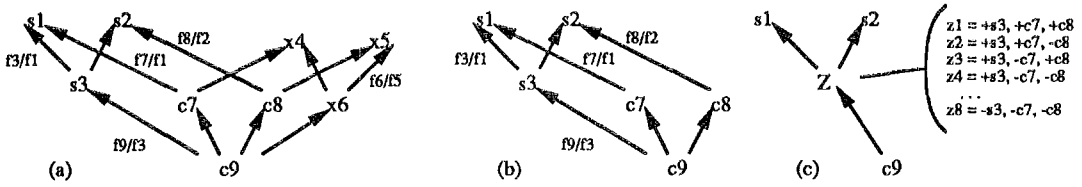

Figure 5. Computing the goodness metric for an "afternoon rest" interpretation (see text).

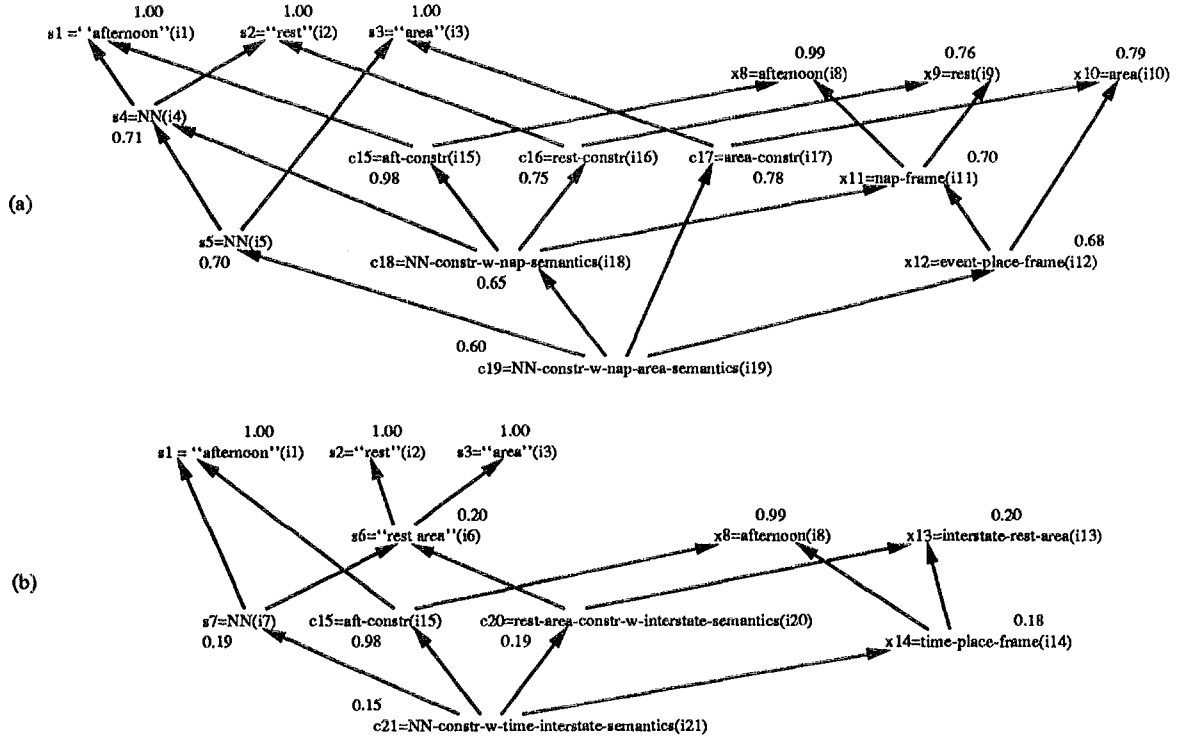

Figure 6. Semantic overriding syntactic preference in "afternoon rest area".

theory to complement the probabilistic model, incorporating both explicit invariant biases and probabilistically learned utility expectations. It is not yet clear whether we shall also need to incorporate pragmatic utility expectations in the constructions.

For methodological reasons we have deliberately impoverished the statistical database, by depriving the model of all information except for category frequencies, relying upon disjunctive interaction to complete the probability model. This limitation on the complexity of statistical information is too restrictive; disjunctive interaction cannot satisfactorily approximate cases where

$$P[+c_3|c_1, c_2] \gg 1 - P[\neg c_3|c_1] \cdot P[\neg c_2|c_1].$$

Such cases appear to arise often; for example, the presence of two nouns, rather than one, increases the probability of a compound by a much greater factor than modeled by disjunctive interaction. We intend to test variants of the model empirically on a corpus of nominal compounds, with randomly selected training sets; the restrictions on complexity of conditional probability information will be relaxed depending upon the resulting prediction accuracy.

Conclusion

We have suggested extending unification-based formalisms to express the sort of interacting preferences used in massively-parallel language models, using probabilistic techniques. In this way, quantitative claims that remain hidden in many massively-parallel models can be made more explicit; moreover, the numbers and the calculus are motivated by a reasonable assumption about language learning. We hope to see increased use of probabilistic models rather than arbitrary calculi in language research: Charniak & Goldman's (1989) recent analysis of probabilities in semantic story structures is a promising development in this direction. Stolcke (1989) transformed a unification grammar into a connectionist framework (albeit without preferences); we have taken the opposite tack. Many linguists have acknowledged the need to extend their frameworks to handle statistically-based syntactic and semantic judgements (e.g., Karlgren 1974; Ford *et al.* 1982, p. 745), but only in passing, largely, we suspect, due to the unavailability of adequate representational tools. Because our proposal makes direct use of traditional unification-based structures, larger grammars should be easy to construct and

incorporate; because of the direct correspondence to semantic net representations, complex semantic models of the type found in AI work may be more readily exploited.

References

Bookman, L. A. (1987). A microfeature based scheme for modelling semantics. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pp. 611–614.

Charniak, E. & R. Goldman (1989). A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1074–1079.

Cottrell, G. W. (1984). A model of lexical access of ambiguous words. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pp. 61–67.

Cottrell, G. W. (1985). A connectionist approach to word sense disambiguation. Technical Report TR 154, Univ. of Rochester, Dept. of Comp. Sci., New York.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, 53(4):810–842.

Fillmore, C. J. (1988). On grammatical constructions. Unpublished draft, University of California at Berkeley.

Ford, M., J. Bresnan, & R. M. Kaplan (1982). A competence-based theory of syntactic closure. In J. Bresnan, editor, *The Mental Representation of Grammatical Relations*, pp. 727–796. MIT Press, Cambridge, MA.

Hirst, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, Cambridge.

Hobbs, J. R., M. Stickel, P. Martin, & D. Edwards (1988). Interpretation as abduction. In *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics*, pp. 95–103, Buffalo, NY.

Jespersen, O. (1946). *A Modern English Grammar on Historical Principles*, volume 6. George Allen & Unwin, London.

Jurafsky, D. S. (1990). Representing and integrating linguistic knowledge. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki.

Karlgren, H. (1974). Categorial grammar calculus. *Statistical Methods In Linguistics*, 1974:1–128.

Lees, R. B. (1963). *The Grammar of English Nominalizations.* Mouton, The Hague.

Leonard, R. (1984). *The Interpretation of English Noun Sequences on the Computer.* North Holland, Amsterdam.

Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

Lytinen, S. L. (1986). Dynamically combining syntax and semantics in natural language processing. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 574–578.

Marcus, M. P. (1980). *A Theory of Syntactic Recognition for Natural Language.* MIT Press, Cambridge.

McClelland, J. L. & A. H. Kawamoto (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In J. L. McClelland & D. E. Rumelhart, editors, *Parallel Distributed Processing*, volume 2, pp. 272–325. MIT Press, Cambridge, MA.

McDonald, D. B. (1982). Understanding noun compounds. Technical Report CMU-CS-82-102, Carnegie-Mellon Univ., Dept. of Comp. Sci., Pittsburgh, PA.

Norvig, P. (1989). Non-disjunctive ambiguity. Unpublished draft, University of California at Berkeley.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

Pollard, C. & I. A. Sag (1987). *Information-Based Syntax and Semantics: Volume 1: Fundamentals.* Center for the Study of Language and Information, Stanford, CA.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (1985). *A Comprehensive Grammer of the English Language.* Longman, New York.

Schubert, L. K. (1986). Are there preference trade-offs in attachment decisions? In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 601–605.

Stolcke, A. (1989). Processing unification-based grammars in a connectionist framework. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 908–915.

Waltz, D. L. & J. B. Pollack (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 9:51–74.

Warren, B. (1978). *Semantic Patterns of Noun-Noun Compounds.* Acta Universitatis Cothoburgensis, Gothenburg, Sweden.

Wermter, S. (1989a). Integration of semantic and syntactic constraints for stuctural noun phrase disambiguation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1486–1491.

Wermter, S. (1989b). Learning semantic relationships in compound nouns with connectionist networks. In *Program of the Eleventh Annual Conference of the Cognitive Science Society*, pp. 964–971.

Wermter, S. & W. G. Lehnert (1989). Noun phrase analysis with connectionist networks. In N. Sharkey & R. Reilly, editors, *Connectionist Approaches to Language Processing.* In press.

Wilensky, R. & Y. Arens (1980). Phran – a knowledge-based approach to natural language analysis. Technical Report UCB/ERL M80/34, University of California at Berkeley, Electronics Research Laboratory, Berkeley, CA.

Wilensky, R., D. Chin, M. Luria, J. Martin, J. Mayfield, & D. Wu (1988). The Berkeley UNIX Consultant project. *Computational Linguistics*, 14(4):35–84.

Wu, D. (1987). Concretion inferences in natural language understanding. In K. Morik, editor, *Proceedings of GWAI-87, 11th German Workshop on Artificial Intelligence*, pp. 74–83, Geseke. Springer-Verlag. Informatik-Fachberichte 152.

Wu, D. (1989). A probabilistic approach to marker propagation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 574–580, Detroit, MI. Morgan Kaufmann.