NOUN CLASSIFICATION FROM PREDICATE-ARGUMENT STRUCTURES

Donald Hindle
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

ABSTRACT

A method of determining the similarity of nouns on the basis of a metric derived from the distribution of subject, verb and object in a large text corpus is described. The resulting quasi-semantic classification of nouns demonstrates the plausibility of the distributional hypothesis, and has potential application to a variety of tasks, including automatic indexing, resolving nominal compounds, and determining the scope of modification.

1. INTRODUCTION

A variety of linguistic relations apply to sets of semantically similar words. For example, modifiers select semantically similar nouns, selectional restrictions are expressed in terms of the semantic class of objects, and semantic type restricts the possibilities for noun compounding. Therefore, it is useful to have a classification of words into semantically similar sets. Standard approaches to classifying nouns, in terms of an "is-a" hierarchy, have proven hard to apply to unrestricted language. Is-a hierarchies are expensive to acquire by hand for anything but highly restricted domains, while attempts to automatically derive these hierarchies from existing dictionaries have been only partially successful (Chodorow, Byrd, and Heidorn 1985).

This paper describes an approach to classifying English words according to the predicate-argument structures they show in a corpus of text. The general idea is straightforward: in any natural language there are restrictions on what words can appear together in the same construction, and in particular, on what can be arguments of what predicates. For nouns, there is a restricted set of verbs that it appears as subject of or object of. For example, *wine* may be *drunk, produced,* and *sold* but not *pruned.* Each noun may therefore be characterized according to the verbs that it occurs with. Nouns may then be grouped according to the extent to which they appear in similar environments.

This basic idea of the distributional foundation of meaning is not new. Harris (1968) makes this "distributional hypothesis" central to his linguistic theory. His claim is that: "the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities." (Harris 1968:12). Sparck Jones (1986) takes a similar view. It is however by no means obvious that the distribution of words will directly provide a useful semantic classification, at least in the absence of considerable human intervention. The work that has been done based on Harris' distributional hypothesis (most notably, the work of the associates of the Linguistic String Project (see for example, Hirschman, Grishman, and Sager 1975)) unfortunately does not provide a direct answer, since the corpora used have been small (tens of thousands of words rather than millions) and the analysis has typically involved considerable intervention by the researchers. The stumbling block to any automatic use of distributional patterns has been that no sufficiently robust syntactic analyzer has been available.

This paper reports an investigation of automatic distributional classification of words in English, using a parser developed for extracting grammatical structures from unrestricted text (Hindle 1983). We propose a particular measure of similarity that is a function of mutual information estimated from text. On the basis of a six million word sample of Associated Press news stories, a classification of nouns was developed according to the predicates they occur with. This purely syntax-based similarity measure shows remarkably plausible semantic relations.

2. ANALYZING THE CORPUS

A 6 million word sample of Associated Press news stories was analyzed, one sentence at a time,

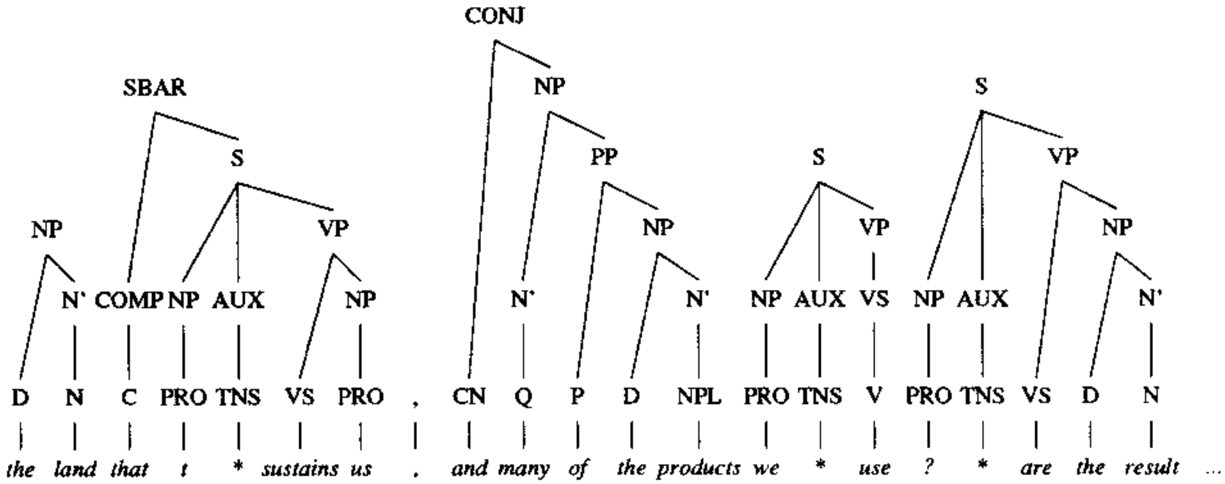
CONJ
SBAR
NP
S
PP
S
S
VP
NP
VP
NP
VP
NP
VP
NP
N' COMP NP AUX
NP
N'
N'
NP AUX VS NP AUX
N'
D N C PRO TNS VS PRO , CN Q P D NPL PRO TNS V PRO TNS VS D N
the land that t * sustains us , and many of the products we * use ? * are the result ...

Figure 1. Parser output for a fragment of sentence (1).

by a deterministic parser (Fidditch) of the sort originated by Marcus (1980). Fidditch provides a single syntactic analysis -- a tree or sequence of trees -- for each sentence; Figure 1 shows part of the output for sentence (1).

(1) *The clothes we wear, the food we eat, the air we breathe, the water we drink, the land that sustains us, and many of the products we use are the result of agricultural research.* (March 22 1987)

The parser aims to be non-committal when it is unsure of an analysis. For example, it is perfectly willing to parse an embedded clause and then leave it unattached. If the object or subject of a clause is not found, Fidditch leaves it empty, as in the last two clauses in Figure 1. This non-committal approach simply reduces the effective size of the sample.

The aim of the parser is to produce an annotated surface structure, building constituents as large as it can, and reconstructing the underlying clause structure when it can. In sentence (1), six clauses are found. Their predicate-argument information may be coded as a table of 5-tuples, consisting of verb, surface subject, surface object, underlying subject, underlying object, as shown in Table 1. In the subject-verb-object table, the root form of the head of phrases is recorded, and the deep subject and object are used when available. (Noun phrases of the form *a n1 of n2* are coded as *n1 n2*; an example is the first entry in Table 2).

Table 1. Predicate-argument relations found in an AP news sentence (1).

verb	**subject**		**object**	
	surface	deep	surface	deep
wear	*we*			
eat	*we*		*0trace*	*food*
breathe	*we*		*0trace*	*air*
drink	*we*		*0trace*	*water*
sustain	*0trace*	*land*	*us*	
use	*we*			
be	*land*		*result*	

The parser's analysis of sentence (1) is far from perfect: the object of *wear* is not found, the object of *use* is not found, and the single element *land* rather than the conjunction of *clothes, food, air, water, land, products* is taken to be the subject of *be*. Despite these errors, the analysis is succeeds in discovering a number of the correct predicate-argument relations. The parsing errors that do occur seem to result, for the current purposes, in the omission of predicate-argument relations, rather than their misidentification. This makes the sample less effective than it might be, but it is not in general misleading. (It may also skew the sample to the extent that the parsing errors are consistent.)

The analysis of the 6 million word 1987 AP sample yields 4789 verbs in 274613 clausal structures, and 26742 head nouns. This table of predicate-argument relations is the basis of our similarity metric.

3. TYPICAL ARGUMENTS

For any of verb in the sample, we can ask what nouns it has as subjects or objects. Table 2 shows the objects of the verb *drink* that occur (more than once) in the sample, in effect giving the answer to the question "what can you drink?"

Table 2. Objects of the verb *drink*.

OBJECT	COUNT	WEIGHT
bunch beer	2	12.34
tea	4	11.75
Pepsi	2	11.75
champagne	4	11.75
liquid	2	10.53
beer	5	10.20
wine	2	9.34
water	7	7.65
anything	3	5.15
much	3	2.54
it	3	1.25
<SOME AMOUNT>	2	1.22

This list of drinkable things is intuitively quite good. The objects in Table 2 are ranked not by raw frequency, but by a cooccurrence score listed in the last column. The idea is that, in ranking the importance of noun-verb associations, we are interested not in the raw frequency of cooccurrence of a predicate and argument, but in their frequency normalized by what we would expect. More is to be learned from the fact that you can *drink wine* than from the fact that you can *drink it* even though there are more clauses in our sample with *it* as an object of *drink* than with *wine*. To capture this intuition, we turn, following Church and Hanks (1989), to "mutual information" (see Fano 1961).

The mutual information of two events $I(x\ y)$ is defined as follows:

$$I(x\ y) \;=\; \log_2 \frac{P(x\ y)}{P(x)\ P(y)}$$

where $P(x\ y)$ is the joint probability of events x and y, and $P(x)$ and $P(y)$ are the respective independent probabilities. When the joint probability $P(x\ y)$ is high relative to the product of the independent probabilities, I is positive; when the joint probability is relatively low, I is negative. We use the observed frequencies to derive a cooccurrence score C_{obj} (an estimate of mutual information) defined as follows.

$$C_{obj}(n\ v) \;=\; \log_2 \frac{\frac{f(n\ v)}{N}}{\frac{f(n)}{N}\ \frac{f(v)}{N}}$$

where $f(n\ v)$ is the frequency of noun n occurring as object of verb v, $f(n)$ is the frequency of the noun n occurring as argument of any verb, $f(v)$ is the frequency of the verb v, and N is the count of clauses in the sample. ($C_{subj}(n\ v)$ is defined analogously.)

Calculating the cooccurrence weight for *drink*, shown in the third column of Table 2, gives us a reasonable ranking of terms, with *it* near the bottom.

Multiple Relationships

For any two nouns in the sample, we can ask what verb contexts they share. The distributional hypothesis is that nouns are similar to the extent that they share contexts. For example, Table 3 shows all the verbs which *wine* and *beer* can be objects of, highlighting the three verbs they have in common. The verb *drink* is the key common factor. There are of course many other objects that can be *sold*, but most of them are less alike than *wine* or *beer* because they can't also be *drunk*. So for example, a *car* is an object that you can *have* and *sell*, like *wine* and *beer*, but you do not -- in this sample (confirming what we know from the meanings of the words) -- typically *drink* a *car*.

4. NOUN SIMILARITY

We propose the following metric of similarity, based on the mutual information of verbs and arguments. Each noun has a set of verbs that it occurs with (either as subject or object), and for each such relationship, there is a mutual information value. For each noun and verb pair, we get two mutual information values, for subject and object,

$$C_{subj}(v_i\ n_j) \quad \text{and} \quad C_{obj}(v_i\ n_j)$$

We define the *object similarity* of two nouns with respect to a verb in terms of the minimum shared coocccurrence weights, as in (2).

The *subject similarity* of two nouns, SIM_{subj}, is defined analogously.

Now define the overall similarity of two nouns as the sum across all verbs of the object similarity and the subject similarity, as in (3).

(2) Object similarity.

$$SIM_{obj}(v_i n_j n_k) = \begin{cases} min(C_{obj}(v_i n_j) \, , C_{obj}(v_i n_k)), & \text{if } C_{obj}(v_i n_j) > 0 \text{ and } C_{obj}(v_i n_k) > 0 \\ abs \; (max(C_{obj}(v_i n_j) \, , C_{obj}(v_i n_k))), & \text{if } C_{obj}(v_i n_j) < 0 \text{ and } C_{obj(}v_i n_k) < 0 \\ 0, & otherwise \end{cases}$$

(3) Noun similarity.

$$SIM(n_1 n_2) = \sum_{i=0}^{N} SIM_{subj}(v_i n_1 n_2) + SIM_{obj}(v_i n_1 n_2)$$

The metric of similarity in (2) and (3) is but one of many that might be explored, but it has some useful properties. Unlike an inner product measure, it is guaranteed that a noun will be most similar to itself. And unlike cosine distance, this metric is roughly proportional to the number of different verb contexts that are shared by two nouns.

Using the definition of similarity in (3), we can begin to explore nouns that show the greatest similarity. Table 4 shows the ten nouns most similar to *boat*, according to our similarity metric. The first column lists the noun which is similar to *boat*. The second column in each table shows the number of instances that the noun appears in a predicate-argument pair (including verb environments not in the list in the fifth column). The third column is the number of distinct verb environments (either subject or object) that the noun occurs in which are shared with the target noun of the table. Thus, *boat* is found in 79 verb environment. Of these, *ship* shares 25 common environments (*ship* also occurs in many other unshared environments). The fourth column is the measure of similarity of the noun with the target noun of the table, $SIM(n_1 n_2)$, as defined above. The fifth column shows the common verb environments, ordered by cooccurrence score, $C(v_i n_j)$, as defined above. An underscore before the verb indicates that it is a subject environment; a following underscore indicates an object environment. In Table 4, we see that *boat* is a subject of *cruise*, and object of *sink*. In the list for *boat*, in column five, *cruise* appears earlier in the list than *carry* because *cruise* has a higher cooccurrence score. A ˜ before a verb means that the cooccurrence score is negative -- i.e. the noun is less likely to occur in that argument context than expected.

Table 3. Verbs taking *wine* and *beer* as objects.

VERB	*wine*		*beer*	
	count	weight	count	weight
drug	2	12.26		
sit around			1	10.29
smell	1	10.07		
contaminate	1	9.75		
rest			2	9.56
drink	2	9.34	5	10.20
rescue			1	7.07
purchase	1	6.79		
lift	1	6.72		
prohibit	1	6.69		
love	1	6.33		
deliver			1	5.82
buy			3	5.44
name	1	5.42		
keep	2	4.86		
offer			1	4.13
begin			1	4.09
allow	1	3.90		
be on			1	3.79
sell	1	4.21	1	3.75
's			2	2.84
make			1	1.27
have	1	0.84	2	1.38

For many nouns, encouragingly appropriate sets of semantically similar nouns are found. Thus, of the ten nouns most similar to *boat* (Table 4), nine are words for vehicles; the most similar noun is the near-synonym *ship*. The ten nouns most similar to *treaty* (*agreement, plan, constitution, contract, proposal, accord, amendment, rule, law, legislation*) seem to make up a cluster involving the notions of *agreement* and *rule*. Table 5 shows the ten nouns most similar to *legislator*, again a fairly coherent set. Of course, not all nouns fall into such neat clusters: Table 6 shows a quite heterogeneous group of nouns similar to *table*, though even here the most similar word (*floor*) is plausible. We need, in further work, to explore both automatic and supervised means of discriminating the semantically relevant associations from the spurious.

Table 4. Nouns similar to *boat*.

Noun	f(n)	verbs	SIM	Verbs
boat	153	79	370.16	_cruise, keel_, _plow, sink_, drift_, step off_, step from_, dock_, right_, submerge_, _near, hoist_, _intercept, charter_, stay on_, buzz_, stabilize_, _sit on, intercept_, hijack_, park_, _be from, rock_, get off_, board_, miss_, stay with_, _catch, yield_, bring in_, seize_, pull_, grab_, _hit, exclude_, weigh_, _issue, demonstrate_, _force, _cover, supply_, _name, _attack, damage_, launch_, _provide, appear_, _carry, _go to, look at_, attack_, _reach, _be on, watch_, use_, return_, _ask, destroy_, fire _, be on_, describe_, charge_, include_, be in_, report_, identify_, expect_, cause_, 's_, _'s, _take, _make, ~be_, ~say_, ~give_, see_, ~_be, ~have_, ~get_
ship	353	25	79.02	_near, charter_, hijack_, get off_, buzz_, intercept_, board_, damage_, sink_, seize_, _carry, attack_, ~have_, _be on, _hit, destroy_, watch_, _go to, ~give_, _ask, ~be_, be on_, ~say_, identify_, see_
plane	445	26	68.85	hijack_, intercept_, charter_, board_, get off_, _near, _attack, _carry, seize_, ~have_, _be on, _catch, destroy_, _hit, be on_, damage_, use_, ~be_, _go to, _reach, ~say_, identify_, _provide, expect_, cause_, see_
bus	104	20	64.49	step off_, hijack_, park_, get off_, board_, _catch, seize_, _carry, attack_, _be on, be on_, charge_, expect_, ~have_, _take, ~say_, _make, include_, be in_, ~_be
jet	153	17	62.77	charter_, intercept_, hijack_, park_, board_, _hit, seize_, _attack, _force, _carry, use_, describe_, include_, _be on, ~_be, _make, ~say_
vessel	172	18	57.14	right_, dock_, _intercept, sink_, seize_, _catch, _attack, _carry, attack_, ~have_, describe_, identify_, use_, report_, ~be_, ~say_, expect_, ~give_
truck	146	21	56.71	park_, intercept_, stay with_, _be from, _hit, seize_, damage_, _carry, _reach, use_, return_, destroy_, attack_, ~_be, be in_, _take, ~have_, ~say_, _make, include_, see_
car	414	24	52.22	step from_, park_, board_, _hit, _catch, pull_, _carry, damage_, destroy_, watch_, miss_, return_, ~give_, ~be_, ~_be, be in_, ~have_, ~say_, charge_, _'s, identify_, see_, _take, ~get_
helicopter	151	14	50.66	hijack_, park_, board_, bring in_, _catch, _attack, watch_, use_, return_, fire_, _be on, include_, _make, ~_be
ferry	37	10	39.76	dock_, sink_, board_, pull_, _carry, use_, be on_, cause_, _take, ~say_
man	1396	30	38.31	hoist_, bring in_, stay with_, _attack, grab_, exclude_, _catch, charge_, ~have_, identify_, describe_, ~give_, _be from, appear_, _go to, _carry, _reach, _take, pull_, _hit, ~get_, 's_, attack_, cause_, _make, ~_be, see_, _cover, _name, _ask

Table 5. Nouns simliar to *legislator*.

Noun	f(n)	verbs	SIM	Verbs
legislator	45	35	165.85	cajole_, _thump, _grasp, convince_, inform_, address_, _vote, _predict, _address, _withdraw, _adopt, _approve, criticize_, _criticize, represent_, _reach, write_, _reject, _accuse, support_, go to_, _consider, _win, pay_, allow_, tell_, _hold, call_, _kill, _call, give_, _get, say_, _take, ~_be
Senate	366	11	40.19	_vote, address_, _approve, inform_, _reject, go to_, _consider, _adopt, tell_, ~_be, give_
committee	697	20	39.97	_vote, _approve, go to_, inform_, _reject, tell_, ~_be, convince_, _hold, address_, _consider, _address, _adopt, call_, _criticize, allow_, support_, _accuse, give_, _call
organization	351	16	34.29	_adopt, inform_, address_, go to_, _predict, support_, _reject, represent_, _call, _approve, ~_be, allow_, _take, say_, _hold, tell_
commission	389	17	34.28	_reject, _vote, criticize_, convince_, inform_, allow_, _accuse, _address, _adopt, ~_be, _hold, _approve, give_, go to_, tell_, _consider, pay_
legislature	86	12	34.12	convince_, _approve, criticize_, _vote, _address, _hold, _consider, ~_be, call_, give_, say_, _take
delegate	132	13	33.65	_vote, inform_, _approve, _adopt, allow_, _reject, _consider, _reach, tell_, give_, ~_be, _call, say_
lawmaker	176	14	32.78	_criticize, _approve, _vote, _predict, tell_, _reject, _accuse, ~_be, call_, give_, _consider, _win, _get, _take
panel	253	12	31.23	_vote, _approve, convince_, tell_, _reject, _adopt, _criticize, _consider, ~_be, _hold, give_, _reach
Congress	827	15	31.20	inform_, _approve, _vote, tell_, _consider, convince_, go to_, ~_be, address_, give_, criticize_, _address, _reach, _adopt, _hold
side	327	15	30.00	_reach, _predict, criticize_, _withdraw, _consider, go to_, _hold, ~_be, _accuse, support_, represent_, tell_, give_, allow_, _take

Table 6. Nouns similar to *table*.

Noun	f(n)	verbs	SIM	Verbs
table	66	30	181.43	hide beneath_, convolute_, memorize_, sit at_, sit across_, redo_, structure_, sit around_, litter_, _carry, lie on_, go from_, _hold, wait_, come to_, return to_, turn_, approach_, cover_, be on_, share_, publish_, claim_, mean_, go to_, raise_, leave_, ~have_, do_, be_
floor	94	6	30.01	litter_, lie on_, cover_, be on_, come to_, go to_
farm	80	8	22.94	_carry, be on_, cover_, return to_, turn_, go to_, leave_, ~have_
scene	135	10	20.85	approach_, return to_, mean_, go to_, be on_, turn_, come to_, leave_, do_, be_
America	156	7	19.68	go from_, come to_, return to_, claim_, go to_, ~have_, do_
experience	129	5	19.04	structure_, share_, claim_, publish_, be_
river	95	4	18.73	sit across_, mean_, be on_, leave_
town	195	6	18.68	litter_, approach_, go to_, return to_, come to_, leave_
side	327	8	18.57	lie on_, be on_, go to_, _hold, ~have_, cover_, leave_, come to_
hospital	190	7	18.10	go from_, come to_, cover_, return to_, go to_, leave_, ~have_
House	453	6	17.84	return to_, claim_, come to_, go to_, cover_, leave_

Reciprocally most similar nouns

We can define "reciprocally most similar" nouns or "reciprocal nearest neighbors" (RNN) as two nouns which are each other's most similar noun. This is a rather stringent definition; under this definition, *boat* and *ship* do not qualify because, while *ship* is the most similar to *boat*, the word most similar to *ship* is not *boat* but *plane* (*boat* is second). For a sample of all the 319 nouns of frequency greater than 100 and less than 200, we asked whether each has a reciprocally most similar noun in the sample. For this sample, 36 had a reciprocal nearest neighbor. These are shown in Table 7 (duplicates are shown only once).

Table 7. A sample of reciprocally nearest neighbors.

RNN			**word counts**
bomb	-	*device*	(192 101)
ruling	-	*decision*	(192 761)
street	-	*road*	(188 145)
protest	-	*strike*	(187 254)
list	-	*field*	(184 104)
debt	-	*deficit*	(183 351)
guerrilla	-	*rebel*	(180 314)
fear	-	*concern*	(176 355)
higher	-	*lower*	(175 78)
freedom	-	*right*	(164 609)
battle	-	*fight*	(163 131)
jet	-	*plane*	(153 445)
shot	-	*bullet*	(152 35)
truck	-	*car*	(146 414)
researcher	-	*scientist*	(142 112)
peace	-	*stability*	(133 64)
property	-	*land*	(132 119)
star	-	*editor*	(131 85)
trend	-	*pattern*	(126 58)
quake	-	*earthquake*	(126 120)
economist	-	*analyst*	(120 318)
remark	-	*comment*	(115 385)
data	-	*information*	(115 505)
explosion	-	*blast*	(115 52)
tie	-	*relation*	(114 251)
protester	-	*demonstrator*	(110 99)
college	-	*school*	(109 380)
radio	-	*IRNA*	(107 18)
2	-	*3*	(105 90)

The list in Table 7 shows quite a good set of substitutable words, many of which are near synonyms. Some are not synonyms but are nevertheless closely related: *economist - analyst, 2 - 3*. Some we recognize as synonyms in news reporting style: *explosion - blast, bomb - device, tie - relation*. And some are hard to interpret. Is the close relation between *star* and *editor* some reflection of news reporters' world view? Is *list* most like *field* because neither one has much meaning by itself?

5. DISCUSSION

Using a similarity metric derived from the distribution of subjects, verbs and objects in a corpus of English text, we have shown the plausibility of deriving semantic relatedness from the distribution of syntactic forms. This demonstration has depended on: 1) the availability of relatively large text corpora; 2) the existence of parsing technology that, despite a large error rate, allows us to find the relevant syntactic relations in unrestricted text; and 3) (most important) the fact that the lexical relations involved in the distribution of words in syntactic structures are an extremely strong linguistic constraint.

A number of issues will have to be confronted to further exploit these structurally-mediated lexical constraints, including:

Polysemy. The analysis presented here does not distinguish among related senses of the (orthographically) same word. Thus, in the table of words similar to *table*, we find at least two distinct senses of *table* conflated; the *table* one can *hide beneath* is not the *table* that can be *commuted* or *memorized*. Means of separating senses need to be developed.

Empty words. Not all nouns are equally contentful. For example, *section* is a general word that can refer to sections of all sorts of things. As a result, the ten words most similar to *section* (*school, building, exchange, book, house, ship, some, headquarter, industry, office*) are a semantically diverse list of words. The reason is clear: *section* is semantically a rather empty word, and the selectional restrictions on its cooccurence depend primarily on its complement. You might read a *section of a book* but not, typically, a *section of a house*. It would be possible to predetermine a set of empty words in advance of analysis, and thus avoid some of the problem presented by empty words. But it is unlikely that the class is well-defined. Rather, we expect that nouns could be ranked, on the basis of their distribution, according to how empty they are; this is a matter for further exploration.

Sample size. The current sample is too small; many words occur too infrequently to be adequately sampled, and it is easy to think of usages that are not represented in the sample. For example, it is quite expected to talk about *brewing beer*, but the pair of *brew* and *beer* does not appear in this sample. Part of the reason for missing selectional pairs is surely the restricted nature of the AP news sublanguage.

Further analysis. The similarity metric proposed here, based on subject-verb-object relations, represents a considerable reduction in the information available in the subjec-verb-object table. This reduction is useful in that it permits, for example, a clustering analysis of the nouns in the sample, and for some purposes (such as demonstrating the plausibility of the distribution-based metric) such clustering is useful. However, it is worth noting that the particular information about, for example, which nouns may be objects of a given verb, should not be discarded, and is in itself useful for analysis of text.

In this study, we have looked only at the lexical relationship between a verb and the head nouns of its subject and object. Obviously, there are many other relationships among words -- for example, adjectival modification or the possibility of particular prepositional adjuncts -- that can be extracted from a corpus and that contribute to our lexical knowledge. It will be useful to extend the analysis presented here to other kinds of relationships, including more complex kinds of verb complementation, noun complementation, and modification both preceding and following the head noun. But in expanding the number of different structural relations noted, it may become less useful to compute a single-dimensional similarity score of the sort proposed in Section 4. Rather, the various lexical relations revealed by parsing a corpus, will be available to be combined in many different ways yet to be explored.

REFERENCES

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd Annual Meeting of the ACL, 299-304.

Church, Kenneth. 1988. A stochastic parts program and noun phrase parser for unrestricted text. Proceedings of the second ACL Conference on Applied Natural Language Processing.

Church, Kenneth and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. Proceedings of the 23rd Annual Meeting of the ACL, 76-83.

Fano, R. 1961. Transmission of Information. Cambridge, Mass:MIT Press.

Harris, Zelig S. 1968. Mathematical Structures of Language. New York: Wiley.

Hindle, Donald. 1983. User manual for Fidditch. Naval Research Laboratory Technical Memorandum #7590-142.

Hirschman, Lynette. 1985. Discovering sublanguage structures, in Grishman, Ralph and Richard Kittredge, eds. Analyzing Language in Restricted Domains, 211-234. Lawrence Erlbaum: Hillsdale, NJ.

Hirschman, Lynette, Ralph Grishman, and Naomi Sager. 1975. Grammatically-based automatic word class formation. Information Processing and Management, 11, 39-57.

Marcus, Mitchell P. 1980. A Theory of Syntactic Recognition for Natural Language. MIT Press.

Sparck Jones, Karen. 1986. Synomyny and Semantic Classification. Edinburgh University Press.