Measuring Verb Similarity

Philip Resnik and Mona Diab
{resnik,mdiab}@umiacs.umd.edu
Department of Linguistics and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD USA

Abstract

The way we model semantic similarity is closely tied to our understanding of linguistic representations. We present several models of semantic similarity, based on differing representational assumptions, and investigate their properties via comparison with human ratings of verb similarity. The results offer insight into the bases for human similarity judgments and provide a testbed for further investigation of the interactions among syntactic properties, semantic structure, and semantic content.

Introduction

The way we model semantic similarity is closely tied to our understanding of how linguistic representations are acquired and used. Some models of similarity, such as Tversky's (1977), assume an explicit set of features over which a similarity measure can be computed, and recent computational methods for measuring word similarity can be thought of as an update of this idea on a large scale, representing words in terms of distributional features acquired via analysis of text corpora (e.g., Brown, Della Pietra, deSouza, Lai, & Mercer, 1992; Schütze, 1993). Other methods, following in the semantic networks tradition of Quillian (1968), focus less on explicit features and more on relationships among lexical items within a conceptual taxonomy, sometimes going beyond taxonomic relationships to also take advantage of frequency information derived from corpora (e.g., Rada, Mili, Bicknell, & Blettner, 1989; Resnik, 1999).

Although some of these approaches are not explicitly designed as cognitive models, we have proposed that prediction of human similarity can provide a useful point of comparison for computational measures of similarity, noting that one must be aware that such comparisons can be quite sensitive to the specific choice of test items (Resnik, 1999). To date, we are only aware of comparisons having been done using noun similarity.

In this paper, we consider the problem of measuring the semantic similarity of verbs. Verb similarity is in many respects a different problem from noun similarity, because verb representations are generally viewed as possessing properties that nouns do not, such as syntactic subcategorization restrictions, selectional preferences, and event structure, and there are dependencies among these properties.[1] This means that particular care must be taken in selecting items, as discussed below, and it also means that the same computational measures may be capturing different properties for verbs than for nouns. For example, the IS-A relationship in WordNet's verb taxonomy (Fellbaum, 1998), central in the computation of some measures, signifies generalization according to manner, as in *devour* IS-A *eat*; concomitantly, the verb taxonomy is considerably wider and shallower than WordNet's noun taxonomy. Similarly, measures based on syntactic dependencies may be sensitive to syntactic adjuncts, such as locative and temporal modifiers, that occur predominantly with verbs rather than with nouns.

In what follows, we first discuss several different measures of word similarity and their properties. We then describe an experiment designed to obtain human similarity ratings for pairs of verbs, discuss the fit of the alternative measures to the human ratings, and suggest some implications of these results for future work.

Models of Verb Similarity

We consider three classes of similarity measure, corresponding to three kinds of lexical representation. In the first, verbs are associated with nodes in a semantic network. In the second, verbs are represented by distributional syntactic co-occurrence features obtained via analysis of a corpus. In the third, verbs are associated with lexical entries represented according to a theory of lexical conceptual structure. These classes of representation can be viewed as occupying three different points on the spectrum from non-syntactic to syntactically relevant facets of verb meaning.

Taxonomic Models

Taxonomic models of lexical and conceptual knowledge have a long history. In this work we use WordNet version 1.5, a large scale taxonomic representation of concepts lexicalized in English. As a model of the lexicon, WordNet's verb hierarchy is limited by design to paradigmatic relations, in explicit contrast to attempts to organize semantically coherent verb classes through shared syntactic behavior.

The simplest and most traditional measure of semantic similarity in a taxonomy counts the number of edges in-

[1] Admittedly, the relevant contrast may turn out not to be part-of-speech *per se*; one could argue that some nouns carry similar kinds of participant information, observing, for example, that *x's gift of y to z* parallels *x gave y to z*. We are not attempting to address that issue here.

tervening between nodes ("edge counting"). A distance in edges is converted to similarity by subtracting from the maximum possible distance in the taxonomy, giving the following measure of distance between verbs w_1 and w_2:

$$\mathrm{wsim}_{\mathrm{edge}}(w_1, w_2) = (2 \times \mathrm{MAX}) - \left[\min_{c_1, c_2} \mathrm{len}(c_1, c_2)\right] \quad (1)$$

where c_1 ranges over $s(w_1)$, c_2 ranges over $s(w_2)$, MAX is the maximum depth of the taxonomy, and $\mathrm{len}(c_1, c_2)$ is the length of the shortest path from c_1 to c_2, with $s(w)$ denoting the set of concepts in the taxonomy that represent senses of word w. If all senses of w_1 and w_2 are in separate sub-taxonomies of the WordNet verb hierarchy their edge-count similarity is defined to be zero.

The simple edge-counting approach has well known problems, and arguments have been made for the following measure of semantic similarity between concepts in a taxonomy based on shared information content (Resnik, 1999):

$$\mathrm{sim}_{\mathrm{info1}}(c_1, c_2) \quad = \quad \max_{c \in S(c_1, c_2)} [-\log \mathrm{p}(c)], \quad (2)$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2, and $-\log \mathrm{p}(c)$ quantifies the "information content" of node c. This yields a measure of verb similarity

$$\mathrm{wsim}_{\mathrm{info1}}(w_1, w_2) \quad = \quad \max_{c1, c2} [\mathrm{sim}_{\mathrm{info1}}(c_1, c_2)], \quad (3)$$

where c_1 ranges over $s(w_1)$ and c_2 ranges over $s(w_2)$, and $\mathrm{p}(c)$ is estimated by observing frequencies in a corpus.[2] Intuitively, the quantity defined in (3) measures the maximum overlap in information between the words being compared. When two words are not very similar, the information content of their most informative subsumer (the node c maximizing $-\log \mathrm{p}(c)$) is low: that subsumer resides high in the taxonomy and thus has high probability, implying low information content. In the most extreme case, the most informative subsumer is just the TOP node of the taxonomy, in which case the probability is 1 and the shared information content (and hence similarity) is 0. When two words are similar, that means there is a node lower in the taxonomy that subsumes them both; being lower in the taxonomy its probability is lower and therefore its information content is higher. Crucially, structural notions such as "lower" and "higher", and the number of intervening arcs between nodes, play no actual role in this model of similarity. As a result, unlike edge counting, this measure does not fall prey to the rampant variation in density within any realistic conceptual taxonomy, where a single IS-A link could represent a tiny semantic distance (e.g. *ballpoint_pen* IS-A *pen*) or a very large semantic distance (e.g. *toy* IS-A *artifact*).[3]

Lin (1998) argues for an alternative information-based measure of similarity that, when applied to a taxonomy, closely resembles the measure just described. It differs in normalizing the shared information content using the sum of the *unshared* information content of each item being compared:

$$\mathrm{sim}_{\mathrm{info2}}(c_1, c_2) \quad = \quad \frac{2 \times \log \mathrm{p}(\bigcap_i C_i)}{\log \mathrm{p}(c_1) + \log \mathrm{p}(c_2)} \quad (4)$$

where the C_i are the "maximally specific superclasses" of both c_1 and c_2. As a result of this normalization, the measure possesses some desirable properties, such as a fixed range from 0 to 1. Word similarity $\mathrm{wsim}_{\mathrm{info2}}$ is defined analogously to Definition (3).

Distributional Co-Occurrence Model

Information-based measures of similarity can be applied to representations other than taxonomic structures. Indeed, Lin demonstrates the generality of the idea by showing how such a measure can be used to measure not only taxonomic distance but also string similarity and the distance between feature sets *à la* Tversky. The latter approach is illustrated by representing words as collections of syntactic co-occurrence features obtained by parsing a corpus. For example, both the noun *duty* and the noun *sanction* would have feature sets containing the feature *subj-of(include)*, but only *sanction* would have the feature *adj-mod(economic)*, since "economic sanctions" appears in the corpus but "economic duties" does not. Because these features include both labeled syntactic relationships and the lexical items filling argument roles, the underlying representational model can be thought of as capturing both syntactic and semantic components of verb meaning.

Lin computes the quantity of shared information as the information in the intersection of the distributional feature sets for the two items being compared. This yields the following measure:

$$\mathrm{wsim}_{\mathrm{distrib}}(w_1, w_2) \quad = \quad \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \quad (5)$$

where $F(w_i)$ is the feature set associated with word w_i, and where $I(\mathcal{S})$, the quantity of information in a feature set $\mathcal{S}$, is computed as $I(\mathcal{S}) = -\sum_{f \in \mathcal{S}} \log \mathrm{p}(f)$.[4] In the experiments described here, we use similarity values obtained for verb pairs using Lin's implementation of his model, with his feature sets and probabilities obtained via analysis of a 22-million-word corpus of newswire text.

Semantic Structure Model

Our third method for assessing the semantic similarity of verbs relies on elaborated representations of verb semantics according to the theory of lexical conceptual structure, or LCS (Dorr, 1993; Jackendoff, 1983). LCS representations make an explicit distinction between *semantic structure*, which characterizes the grammatically relevant facets of verb meaning, from *semantic content*, which characterizes idiosyncratic information associated with the verb but not reflected in its syntactic behavior.

[2] For taxonomic measures described in this section, probabilities of nodes in WordNet 1.5 were estimated on the basis of word frequencies in the Brown Corpus (Francis & Kučera, 1982).

[3] Examples are from WordNet 1.5, where *artifact* signifies a man-made object.

[4] Note the assumption that features are independent, permitting the summation of log probabilities.

This difference between semantic structure and semantic content plays an important role in current research on lexical representation (e.g. Grimshaw, 1993; Pinker, 1989; Rappaport, Laughren, & Levin, 1993). We take advantage of this distinction here to derive a measure that focuses exclusively on similarity of semantic structure as disentangled from semantic content.

To illustrate with a simple example, within an LCS representational system *roll* and *slide* might both have semantic structure indicating a change of location, e.g.,

$$
\begin{array}{l}
(\mathrm{go}_{\mathrm{loc}}\ \mathrm{x} \\
\quad (\mathrm{to}_{\mathrm{loc}}\ \mathrm{x}\ (\mathrm{at}_{\mathrm{loc}}\ \mathrm{x}\ \mathrm{y})) \\
\quad (\mathrm{from}_{\mathrm{loc}}\ \mathrm{x}\ (\mathrm{at}_{\mathrm{loc}}\ \mathrm{x}\ \mathrm{z})) \\
\quad (\mathrm{manner}\ \langle \mathrm{M} \rangle)),
\end{array}
$$

and differ only in the value $\langle \mathrm{M} \rangle$ — an element of semantic content within the semantic structure — indicating the manner of motion (either $\langle$SLIDING$\rangle$ or $\langle$ROLLING$\rangle$). Such regularities in semantic structure are argued to provide an explanation for systematic relationships between meaning and syntactic realization (Levin & Rappaport Hovav, 1998).

If those regularities are a part of verb lexical representations, then they also plausibly influence ratings of verb similarity, and the question is how to assess similarity between two such structured representations. Lin's work provides one plausible answer: decomposing complex representations into (pseudo-)independent feature sets and then comparing feature sets.[5] Our method of decomposition was particularly simple, recursively creating an independent feature from each primitive component of the representation and the "head" of its subordinates. So, for example, the feature set representation of *roll* would contain six features:

$$
\begin{array}{l}
[\mathrm{go}_{\mathrm{loc}}\ \mathrm{to}_{\mathrm{loc}}\ \mathrm{from}_{\mathrm{loc}}\ \mathrm{manner}] \\
[\mathrm{to}_{\mathrm{loc}}\ \mathrm{x}\ \mathrm{at}_{\mathrm{loc}}] \\
[\mathrm{at}_{\mathrm{loc}}\ \mathrm{x}\ \mathrm{y}] \\
[\mathrm{from}_{\mathrm{loc}}\ \mathrm{x}\ \mathrm{at}_{\mathrm{loc}}] \\
[\mathrm{at}_{\mathrm{loc}}\ \mathrm{x}\ \mathrm{z}] \\
[\mathrm{manner}\ \langle \mathrm{ROLLING} \rangle].
\end{array}
$$

The features of *slide* would be identical but for the last feature, which would instead be [manner $\langle$SLIDING$\rangle$], and the nearly complete overlap between the feature sets for the two verbs captures the fact that the semantic distinction between this particular pair of verbs rests entirely on semantic content and not semantic structure.

Since we had available to us a large lexicon of LCS representations for verbs in English (Dorr & Olsen, 1996, 1997), containing thousands of lexical entries, we estimated the probability of each feature by counting feature occurrences within the lexicon. We define the similarity of two LCS lexicon entries e_1 and e_2 using the shared information content of their feature sets:

$$\mathrm{sim}_{\mathrm{lcs}}(e_1, e_2) \;=\; I(F(e_1) \cap F(e_2)) \qquad (6)$$

using $I(S)$ as in (5), and we compute $\mathrm{wsim}_{\mathrm{lcs}}(w_1, w_2)$ as the maximum value of $\mathrm{sim}_{\mathrm{lcs}}$ taken over the cross product of all the words' lexical entries.[6]

It is worth emphasizing that this similarity measure considers *only* semantic structure, not semantic content, and therefore only syntactically relevant components of meaning enter into the computation. For example, in the comparison of LCS entries for *slide* and *roll*, $F(e_1) \cap F(e_2)$ will never contain either [manner $\langle$ROLLING$\rangle$] or [manner $\langle$SLIDING$\rangle$], and therefore any potential similarities or differences between the content elements — the *physical* aspects of sliding motion versus rolling motion based on real-world knowledge — are excluded from the model.

Experiment

In order to assess alternative computational models of similarity, we collected human ratings of similarity for pairs of verbs, following a design after that of Miller and Charles (1991). Considering the additional complexities in the verb lexicon, however, the selection of materials required considerable care: we were careful to pay close attention to syntactic subcategorization, thematic grids, and aspectual class information, as described below, in order to limit the possible dimensions across which the two verbs in a pair could differ and to focus on *semantic* similarity. We also designed two versions of the task, with and without presentation of verbs in context, in order to investigate the extent to which contextual narrowing of verbs' senses affects ratings of similarity.

Participants. Participants were 10 volunteers, all native speakers of English, ranging in age from 24 to 53, without significant background in psychology or linguistics. All participated by e-mail.

Materials. In constructing the set of verb pairs for similarity ratings, we began with the set of verbs in a large lexicon of LCS entries, containing entries for 4900 verbs. Verb entries in the lexicon contain information about both aspectual features (dynamicity, durativity, telicity; Olsen, 1997) and thematic grid (identifying whether or not a verb takes an agent, theme, goal, etc.) — for example, the verb *broil* requires both an agent and a theme, and is marked as both durative and telic but not dynamic. For subcategorization information, we referred to the Collins Cobuild dictionary (Sinclair, 1995), using the subcategorization frame for the first listed verb sense.

To construct verb pairs, we began by eliminating all verbs whose thematic grid did not require a theme, in order to limit the range of variation in thematic grids.[7]

[5] We are grateful to Dekang Lin for suggesting this approach to us.

[6] Although our probability estimate counts features within a set of types (entries in a large lexicon) rather than tokens (verb instances in a large corpus), inspection of the estimated probabilities suggests that frequent features are suitably discounted, having low information content, and rare features are highly informative. Corpus-based estimates are a matter for future work.

[7] All verbs require an agent, so the remaining variation is in the presence or absence of oblique roles such as GOAL.

We then grouped the full set of verbs into eight lists corresponding to the eight possible combinations of the three aspectual features, and restricted our attention to the four most numerous lists.[8] Within each of those four lists, we created 12 pairs of verbs subject to the constraint that the verbs' associated subcategorization frames had to match, so as to avoid effects of purely syntactic similarity. Items were selected to span the range from low- to high-similarity verb pairs.

In summary, a set of 48 verb pairs was constructed so that (i) both verbs in every pair require a theme, (ii) both verbs have the same subcategorization frame, and (iii) both verbs come from the same aspectual class. Verbs on the list were all given in the past tense. In order to avoid ordering effects, half the subjects in each condition saw items in a random order, and the other half saw the items in the reverse order.

To assess the effects that contextual narrowing of verb senses might have on similarity ratings, the materials as just described were duplicated in order to create *No Context* and *Context* conditions. The conditions were identical except that in the *Context* condition, each item was accompanied by an example sentence for each verb illustrating the verb's intended sense. Each example sentence came from the corresponding verb entry in the Collins Cobuild dictionary. For example, the example sentence for *loosen* was "He loosened his seat belt."

Procedure. The 10 subjects were split evenly into *Context* and *No Context* groups. Subjects in the *No Context* group were given the set of 48 verb pairs, without example sentences, and asked to compare their meanings on a scale of 0–5, where 0 means that the verbs are not similar at all and 5 indicates maximum similarity. Subjects were explicitly asked to ignore similarities in the sound of the verb and similarities in the number and type of letters that make up the verb. Subjects were also asked explicitly to rate similarity rather than relatedness, with the instructions giving an example of the distinction. (For example, *pay* and *eat* are related in that they are things we do in restaurants, but they are not particularly similar.) Since some verbs in the set have low frequency, a "don't know" box was included for subjects to mark if they were unsure of the meaning of either verb. There was no time limit on the task, which tended to take approximately 20 minutes.

Subjects in the *Context* group were given exactly the same task, but using the *Context* materials, i.e. with each verb accompanied by an example sentence illustrating the intended sense. As in the previous condition, two orders of presentation were used within this condition to avoid ordering effects.

Each computational similarity measure took the set of verb pairs as input, without context, and computed a similarity score for each.

[8] These were {durative}, {durative,dynamic}, {dynamic,telic}, {durative,dynamic,telic}. Verbs could and did appear on multiple lists.

Table 1: Comparing sets of ratings

wsim	*Context*	*No Context*
edge	.720	.675
info1	.779	.658
info2	.768	.668
distrib	.453	.433
lcs	.313	.385
Combined	.872	.785
Inter-rater	.793	.764

Results and Discussion. In order to judge the degree to which sets of similarity ratings are predictive of each other, we use a similarity coefficient computed as Pearson's r. Table 1 provides a summary showing r for each computational model as compared to the mean of the human subject ratings in the *Context* and *No Context* conditions.[9]

The *Combined* row of the table shows the value of multiple R when the five computational measures are compared with human ratings using a multiple regression (see below), and the *Inter-rater* row of the table shows human average inter-rater agreement, measured by r, using leave-one-out resampling (Weiss & Kulikowski, 1991).

Examining these figures, we first consider each computational model separately. It is unsurprising that the similarity measure based on LCS representations fares worst, given the design of the experiment: the verb pairs were selected so as to eliminate differences of subcategorization frame, aspectual class, and thematic grid, ruling out *a priori* pairs that differ interestingly with respect to semantic structure. The distributional measure based on syntactic co-occurrence features may be a victim of its dependence on a particular corpus, and of data sparseness — for example, glaring divergences with human ratings include some verb pairs containing some lower-frequency words, such as *embellish/decorate* and *dissolve/dissipate*. Turning to the taxonomic methods, the information-based approaches appear superior to edge counting in the *Context* condition, consistent with previous work on noun similarity, though in the *No Context* condition there are no clear differences. We suspect a difference will emerge with a larger set of items, but this remains to be seen. Our inspection of by-item

[9] From the full set of items, 10 verb pairs were excluded because some participant did not know the meaning of one or the other verb. Moreover, in preparation of the final version of this paper, we discovered that 11 verb pairs inadvertently had been included despite failing to strictly match the criteria described in the Materials section or having other minor errors of presentation, and these are now excluded, as well. Although this is a large number of excluded items, we consider them quite unlikely to have affected participants' judgments since the excluded pairs were distributed almost perfectly evenly over the four verb lists and varied across degrees of similarity, and since the pattern of results was unaffected. We report all quantitative results in the paper based on only the 27 non-excluded verb pairs.

ratings of the information measures suggests strongly that the differences between the unnormalized and normalized information-based measures are small in comparison to the role played by the structure of the WordNet verb taxonomy.

Comparison of human raters yields several interesting observations. First, a comparison of the *Context* and *No Context* mean ratings by human participants yields $r = .89$, which provides some reassurance that subjects in the *No Context* condition are generally interpreting the verbs in the same sense as are subjects in the *Context* condition — where, recall, the context sentence encouraged interpretation according to the first listed verb sense in the Collins Cobuild dictionary. Second, however, average inter-rater agreement in the two conditions (.79 and .76) is much lower than that obtained in a noun ratings experiment using the same method, where leave-one-out resampling yielded an estimate of $r = .90$ (Resnik, 1999). This may reflect the small sample size in each group ($N = 5$), but we suspect that in actuality it is evidence that word similarity is harder for subjects to quantify for verbs than for nouns. Third, we find that subjects in the *No Context* condition have a very strong tendency to assign higher similarity ratings to the same pair as compared to subjects in the *Context* condition, as determined using a paired t-test ($N = 27, t(26) = 4.49, p < .0002$).

This last observation is consistent with the idea that subjects in the *No Context* condition are accommodating verb comparisons — allowing for more flexible interpretations of verb meaning — in a way not available to subjects in the *Context* condition because their interpretations are constrained by the context sentence. For example, the verb pair *compose/manufacture* has a mean rating of 2.8 in the *Context* condition, and the context sentences are *He sees the whole, not the various lines that compose it* and *Many factories were manufacturing desk calculators.* In the *No Context* condition, the mean rating for this pair is 4.0, likely indicating that in the process of comparison, subjects focused on available semantic elements of *compose*'s meaning that are closest to *manufacture* (e.g., the notion of composing as creating, *She composed satirical poems for the New Statesman*).

As a preliminary step toward combining models, we performed a multiple regression predicting human ratings using the ratings of the five computational models as independent variables, with the results shown in Table 1 as *Combined.* Although we have not extensively analyzed these data, regressions using all $2^5 - 1 = 31$ combinations of models show that the highest multiple R is obtained when all five models are combined, that the two different information-based measures are making essentially the same contribution to the combined model (consistent with our observation that WordNet structure plays the dominant role, rather than details of the measure), and that the LCS measure contributes little for this set of items. Taking these observations into account, the improvement in predictive power when combining models comes from distributional and information-based models being sensitive to at least some different information.

General Discussion

The experimental results reflect the fact that similarity measures model different aspects of verb representation and use. Taxonomic similarity measures place little emphasis on verbs' argument structure, emphasizing relationships of semantic content; for example, *drag* and *tug* appear quite close in the taxonomy (under *displace*) although they differ significantly in semantic structure (e.g. in "the tailpipe dragged" and "the donkey tugged" the syntactic subjects have different thematic roles). Conversely, semantic structure is emphasized in the measure based on LCS representations to the exclusion of real-world knowledge, such as the similarity of the physical motions of dragging and tugging. Distributional similarity based on syntactic co-occurrence features is a combination, capturing elements of semantic structure by means of the syntactic relationships (one-versus two-participant relationships), and also indirectly capturing elements of semantic content by means of the lexical items co-occurring in those syntactic positions (*tug* being weighted more heavily against inanimate subjects than *drag*, for example). Based on the performance of the models, and improved predictive power of the multiple regression, we interpret our results as evidence that human ratings of similarity are sensitive to both paradigmatic and syntagmatic facets of verb representation, and we believe the computational models are capturing relevant aspects of verb representation in order to make predictions about similarity judgments.

On a somewhat speculative note, it is interesting to briefly examine cases where the computational models fail to capture similarities identified by the human raters. Consider, for example, items *unfold/divorce*, *chill/toughen*, *initiate/enter*. Based on the WordNet taxonomy, the verbs in these pairs have no common subsumer, so the shared information content is zero; nor do the distributional or LCS measures predict that they are at all similar. The human mean ratings are low (averaging 1.6, 1.4, and 3.2, respectively, in the *No Context* condition), but why are they not zero — and why are they in fact higher than the ratings for some other pairs, such as *open/inflate* (0.6), where one could also identify reasons for believing the meanings have something in common? It would appear that in these cases subjects are finding similarities of meaning according to dimensions that we have not yet formalized. The apparent sense extensions verge on the metaphorical: one can describe divorce as the unfolding of a marriage, observe a person chill and toughen in response to an insult, enter a group by being initiated into it. Capturing those dimensions of similarity in our models will require a better understanding than we have at present of how word meanings are represented and organized.

Even for the time being, however, the work described in this paper offers a method and a testbed for investigating lexical issues that can go well beyond the present experiments. We chose here to tightly control aspect and syntactic subcategorization while allowing our test items to differ on thematic grids and vary widely with respect to semantic content. Having validated the approach — performance being consistent with what one would predict of the alternative models given the design of the task — the initial work opens the door to other configurations, controlling variation among subcategorization frames, aspectual features, thematic grids, and semantic content in other combinations. What is crucial is that implemented models of similarity, drawing on such theoretical constructs, yield testable predictions that can be verified through careful experimentation.

Acknowledgments

We are grateful to Dekang Lin and Amy Weinberg for valuable discussions, to Dekang Lin for his kindly computing values of distributional similarity (Definition 4) for the verb pairs in our experiment, and to three anonymous reviewers for their helpful comments. This work was supported in part by DARPA/ITO Contract N66001-97-C-8540.

Appendix: Verb Pairs

bathe	kneel	loosen	open
chill	toughen	neutralize	energize
compose	manufacture	obsess	disillusion
compress	unionize	open	inflate
crinkle	boggle	percolate	unionize
displease	disillusion	plunge	bathe
dissolve	dissipate	prick	compose
embellish	decorate	swagger	waddle
festoon	decorate	unfold	divorce
fill	inject	wash	sap
hack	unfold	weave	enrich
initiate	enter	whisk	deflate
lean	kneel	wiggle	rotate
loosen	inflate		

References

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–480.

Dorr, B. J. (1993). *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA.

Dorr, B. J., & Olsen, M. B. (1996). Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization. *Machine Translation*, *11*(1–3), 37–74.

Dorr, B. J., & Olsen, M. B. (1997). Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pp. 151–158 Madrid, Spain.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database.* MIT Press.

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, Boston.

Grimshaw, J. (1993). Semantic structure and semantic content in lexical representation. unpublished manuscript, Center for Cognitive Science, Rutgers University, New Brunswick, New Jersey.

Jackendoff, R. (1983). *Semantics and Cognition.* The MIT Press, Cambridge, MA.

Levin, B., & Rappaport Hovav, M. (1998). Building Verb Meanings. In Butt, M., & Geuder, W. (Eds.), *The Projection of Arguments: Lexical and Compositional Factors*, pp. 97–134. CSLI Publications, Stanford, CA.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98)* Madison, Wisconsin.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Olsen, M. B. (1997). *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect.* Garland, New York.

Pinker, S. (1989). *Learnability and Cognition.* MIT Press, Cambridge, MA.

Quillian, M. R. (1968). Semantic memory. In Minsky, M. (Ed.), *Semantic Information Processing.* MIT Press, Cambridge, MA.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, *19*(1), 17–30.

Rappaport, M., Laughren, M., & Levin, B. (1993). Levels of lexical representation. In Pustejovsky, J. (Ed.), *Semantics and the Lexicon.* Kluwer.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, *11*, 95–130. http://www.cs.washington.edu/research/jair/abstracts/resnik99a.html

Schütze, H. (1993). Word space. In Hanson, S. J., Cowan, J. D., & Giles, C. L. (Eds.), *Advances in Neural Information Processing Systems 5*, pp. 895–902. Morgan Kaufmann Publishers, San Mateo CA.

Sinclair, J. (Ed.). (1995). *Collins Cobuild English Dictionary.* Collins. Patrick Hanks, managing editor.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* Morgan Kaufmann, San Mateo, CA.