A Preliminary Model of Pronoun/Verb Co-occurrences in Child-Directed Speech

Aarre Laakso (alaakso@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Department of Psychology, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

Given the selectional restrictions on the kinds of subjects and objects that a verb may take, it seems likely that children learn verbs partly by exploiting statistical regularities in co-occurrences between verbs and nouns. This paper explores the role of pronouns in this process. Although pronouns are semantically "light," they dominate nouns in the speech children hear and systematically partition important classes of verbs. We show that a statistical learner can exploit these regularities to constrain the possible verbs that might fit in a simple syntactic frame.

Introduction

Pronouns stand for central elements of adult conceptual schemes—as Quine pointed out, pronouns "are the basic media of reference" (Quine, 1980, p. 13). Most syntactic subjects in spontaneous spoken adult discourse are pronouns (Chafe, 1994), and preferred argument structure analyses have suggested that, for information processing reasons, pronouns *should* be the most common subjects of transitive and copular clauses, because they maintain the topic (DuBois, Kumpf, & Ashby, 2003). When addressing their children, English-speaking mothers often begin with a high-frequency pronoun — *you* and *I* occur most frequently (Valian, 1991). Parents use the inanimate pronoun *it* far more frequently as the subject of an intransitive sentence than of a transitive one (Cameron-Faulkner, Lieven, & Tomasello, 2003). As Cameron-Faulkner et al. note, this suggests that intransitive sentences are used more often than transitives for talking about inanimate objects. It also suggests, we would note, that the use of the inanimate pronoun might be a cue for the child as to some aspects of the verb (i.e., whether it is transitive or intransitive). Similarly, Childers & Tomasello (2001) have suggested that pronouns may form the fixed element in lexically-specific frames acquired by early language learners—so-to-speak "pronoun islands," something like Tomasello's (1992) "verb islands."

What has not been studied is how these "basic media of reference" may help children learn other words by virtue of systematic co-occurrences. We address this issue by measuring the statistical regularities among the uses of pronouns and verbs in a large corpus of parent and child speech and by modeling them with a connectionist network.

Experiment 1

The first experiment consisted of a corpus analysis to identify patterns of co-occurrence between pronouns and verbs in the child's input.

Method

Parental utterances from the CHILDES database (MacWhinney, 2000) were coded for syntactic categories, then subjected to cluster analysis. The target children in the transcripts were aged approximately 1;4 – 6;1.

Materials The following corpora were used: Bates, Bliss, Bloom 1970, Brown, Clark, Cornell, Demetras Working, Gleason, Hall, Higginson, Kuczaj, MacWhinney, Morisset, New England, Post, Sachs, Suppes, Tardiff, Valian, Van Houten, Van Kleeck and Warren-Leubecker.[1] Coding was performed using an Internet application that randomly selected transcripts, assigned them to coders as they became available, collected coding input, and stored it in a MySQL database. The application occasionally assigned the same transcript to all coders, in order to measure reliability. Five undergraduate coders were trained on the coding task and the use of the system.

Procedure Each main tier line was coded for speaker, addressee, and syntactic frame (no verb, question, passive, copula, intransitive, transitive or ditransitive). Each word was then coded for its syntactic category in that utterance (subject, auxiliary, verb, direct object, indirect object and oblique — others were ignored). In total, 59,977 utterances were coded from 123 transcripts. *All* of the coders coded 7 of those transcripts for the purpose of measuring reliability. Average inter-coder reliability (measured for each coder as the percentage of items coded exactly the same way they were coded by each other coder) was 86.1%.

We only considered parental child-directed speech (PCDS), defined as utterances where the speaker was a parent and the addressee was a target child. Clauses with no verb, questions, passives and copulas were excluded from further analysis; thus, the analysis was conducted using only

[1] Citations to original sources for the corpora are omitted here due to space constraints. The full references for each corpus may be found in (MacWhinney, 2000).

clauses that were intransitives, transitives, or ditransitives, a total of 12,377 clauses.

From these clauses, we formed 2 matrices: a verbs-by-subjects matrix and a verbs-by-objects matrix. The verbs-by-subjects matrix contained only verbs used with an overt subject; its size was 621 verbs by 317 nouns (subjects). The verbs-by-objects matrix contained only verbs used with a direct object; its size was 524 verbs by 907 nouns (objects). Each cell of each matrix contained the proportion of times that verb was used with that noun (as subject or object) in a coded clause.

We then performed 4 cluster analyses. First, we took the 50 nouns most commonly used as objects and clustered them according to their proximity in verb space, i.e., the space formed by considering each verb as a dimension. Each noun was placed along each dimension according to the proportion of times it was used with the corresponding verb. Hence, a noun never used as the object of a particular verb would be at 0, and a noun always used as the object of a particular verb would be at 1. Second, we clustered the 50 most common subject-nouns in verb space. Third, we took the 50 verbs most commonly used with objects and clustered them according to their proximity in noun space (defined analogously to verb space). Finally, we clustered the 50 most common verbs-with-subjects in noun space. As another means of understanding the structure of the data, for each of the four cluster diagrams, we also plotted the corresponding words in the principal components of the relevant space.

Results

We cannot show all of the cluster diagrams or principal components plots here due to space limits. However, the plot of verbs in the principal components of the space of syntactic objects is shown in Figure 4 on page 6 below, and the corresponding cluster diagram is shown in Figure 5 on page 6 below.

Overview The most frequent nouns in the corpus — both subjects and objects — are pronouns, as shown in Figures 1 and 2. The objects divided the most common verbs into three main classes: verbs that take the pronoun *it* and concrete nouns as objects, verbs that take complement clauses, and verbs that take specific concrete nouns as objects. This may be observed in Figures 4 and 5 — approximately the lower 1/3 of the verbs in Figure 5 take complement clauses instead of nominal objects, whereas the middle 1/3 take *it* as an object, and the upper 1/3 take primarily concrete nouns. This may also be observed in Figure 4, in which the verbs that take clauses as objects are clumped in the lower right, the verbs that take *it* as an object are clumped in the lower left, and the verbs that take a variety of concrete nouns are scattered across the rest of the figure. The subjects divided the most common verbs into four main classes: verbs whose subject is almost always *I*, verbs whose subject is almost always *you*, verbs whose subject is almost always either *I* or *you*, and other verbs. The verbs divided the most common object nouns into a number of classes, including objects of telling and looking verbs, objects of having and wanting verbs, and objects of putting and getting verbs. The verbs also divided the most common subject nouns into a number of classes, including subjects of having and wanting verbs, and subjects of thinking and knowing verbs.

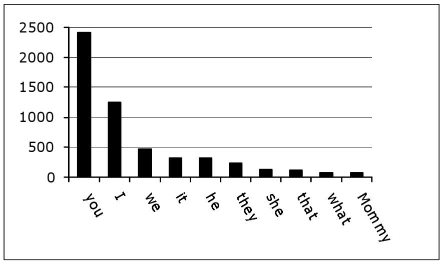

Figure 1: The 10 most frequent subjects in PCDS by their number of occurrences

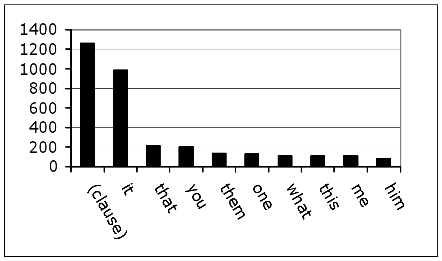

Figure 2: The 10 most frequent objects in PCDS by their number of occurrences.

Verbs that take *it* as an object The verbs that take *it* as their most common object include verbs of motion and transfer, as shown in Table 1.

Table 1: Verbs most commonly used with object *it*.

Verb	Total	it (#)	it (%)
turn	56	33	58.9
throw	36	20	55.5
push	25	13	52.0
hold	42	19	45.2
break	36	16	44.4
leave	27	12	44.4
open	36	15	41.7
do	256	105	41.0
wear	25	10	40.0
take off	24	9	37.5
put	276	93	33.7
get	348	74	21.3
take	106	22	20.8
put on	42	8	19.0
buy	50	9	18.0
give	85	14	16.5
have	340	26	7.6

Verbs that take complement clauses Most verbs that did not take *it* as their most common object instead took complement clauses. These are primarily psychological verbs, as shown in Table 2.

Verbs that take *I* as a subject Verbs whose most common subject is *I* include *bet* (23 out of 23 uses with a subject, or 100%), *guess* (21/22, 95.4%), *think* (212/263, 80.6%), and *see* (95/207, 45.9%). Parents were not discussing their gambling habits with their children — *bet* was being used to indicate the epistemic status of a subsequent clause, as were the other verbs.

Table 2: Verbs most used most with complement clauses.

Verb	Total	<clause> (#)	<clause> (%)
think	187	179	95.7
remember	31	23	74.2
let	78	57	73.1
know	207	141	68.1
ask	29	17	58.6
go	55	32	58.2
want	317	183	57.7
mean	25	14	56.0
tell	115	45	39.1
try	51	18	35.3
say	175	53	30.3
look	48	14	29.2
need	64	18	28.1
see	266	73	27.4
like	123	32	26.0
show	36	9	25.0
make	155	23	14.8

Table 3: Some verbs commonly used with subject *I* or *you.*

Verb	Total	I (#)	I (%)	you (#)	you (%)
bet	23	23	100	0	0
guess	22	21	95.4	0	0
think	263	212	80.6	38	14.4
see	207	95	45.9	50	24.1
mean	32	15	46.9	12	37.5
know	360	159	44.2	189	52.5
remember	23	9	39.1	12	52.2
like	134	20	14.9	86	64.2
want	270	34	12.6	192	71.1
need	65	5	7.7	33	50.8

Verbs that take *you* as a subject Verbs whose most common subject is *you* include *like* (86 out of its 134 total uses with a subject, or 64.2%), *want* (192/270, 71.1%), and *need* (33/65, 50.8%). These verbs are being used to indicate the deontic status of a subsequent clause, i.e., the subject's disposition or inclination, volition, or compulsion with respect to the proposition expressed by the complement.

Verbs that take *you* or *I* as a subject Verbs that take *I* and *you* more or less equally as subject include *mean* (15 out of 32 uses, or 46.9%, with *I* and 12 of 32 uses, or 37.5%, with *you*), *know* (*I*: 159/360, 44.2%; *you*: 189/360, 52.5%), and *remember* (*I*: 9/23, 39.1%; *you*: 12/23, 52.2%).

Subjects of *think* and *know* The subject *I* appeared most frequently with the verbs *think* and *know.*

Discussion

Although pronouns are semantically "light," their particular referents determinable only from context, they may nonetheless be potent forces on early lexical learning by statistically pointing to some classes of verbs as being more likely than others. The results of Experiment 1 clearly show that there are statistical regularities in the co-occurrences of pronouns and verbs that discriminate between broad classes of verbs. The verb clusters identified in Experiment 1 share more than their associations with pronouns — each cluster corresponds roughly to a broad class of verbs with similar semantic aspects. Specifically, when followed by *it*, the verb is likely to describe physical motion, transfer, or possession. When followed by a relatively complex complement clause, by contrast, the verb is likely to attribute a psychological state. If the subject is *I*, the verb is likely to have to do with thinking or knowing, whereas if the subject is *you*, *she*, *we*, *he*, or *they*, the verb is likely to have to do with having or wanting. This regularity most likely reflects the ecology of parents and children — parents "know" and children "want" — but it could nonetheless be useful in distinguishing these two classes of verbs.

The results thus far show that there are potentially usable regularities in the statistical relations between pronouns and verbs. However, they do not show that these regularities can be used to cue the associated words.

Experiment 2

To demonstrate that the regularities in pronoun-verb co-occurrences in parental speech to children can actually be exploited by a statistical learner, we trained a connectionist network to auto-associate subject-verb-object "sentences" from the input, then tested it on individual verbs and pronouns.

Method

Data The network training data consisted of the subject, verb, and object of all coded utterances that contained the 50 most common subjects, verbs and objects. There were 5,835 such utterances. The inputs used a localist coding wherein there was one and only one input unit out of 50 activated for each subject, and likewise for each verb and each object. Absent and omitted arguments were counted among the 50, so, for example, the utterance "John runs" would have 3 units activated even though it only has 2 words — the third unit being the "no object" unit. With 50 units each for subject, verb and object, there were a total of 150 input units to the network. Active input units had a value of 1, and inactive input units had a value of 0.

Network Architecture The network consisted of a two-layer 150-8-150 unit autoassociator with a logistic activation function at the hidden layer and three separate softmax activation functions (one each for the subject, verb and object) at the output layer—see Figure 3. Using the softmax activation function, which ensures that all the outputs in the bank sum to 1, together with the cross-entropy error measure, allows us to interpret the network outputs as probabilities (Bishop, 1995). The network was trained by the resilient backpropagation algorithm (Riedmiller & Braun, 1993) to map its inputs back onto its outputs. It is well known that this sort of network performs nonlinear dimensionality reduction at its hidden layers, extracting statistical regularities from the input data.

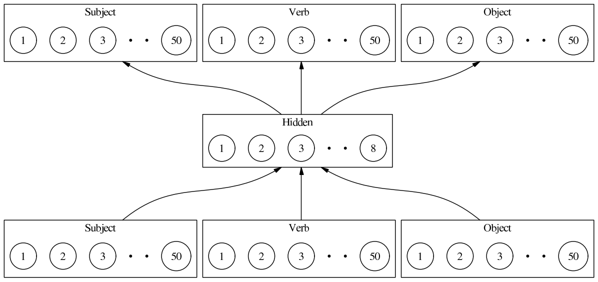

Figure 3: Network architecture

Training The data was randomly assigned to two groups: 90% of the data was used for training the network, while 10% was reserved for validating the network's performance. Starting from different random initial weights, five networks were trained until the cross-entropy on the validation set reached a minimum for each of them. (Multiple networks were used in order to ensure that the results were not idiosyncratic.) Training stopped after approximately 150 epochs of training, on average. At that point, the networks were achieving about 81% accuracy on correctly identifying subjects, verbs and objects from the training set. Near perfect accuracy on the training set could have been achieved by further training, with some loss of generalization. We decided that it was better to avoid over fitting.

Testing After training, the networks were tested with incomplete inputs corresponding to isolated verbs and pronouns. For example, to see what a network had learned about *it* as a subject, it was tested with a single input unit activated — the one corresponding to *it* as subject. The other input units were set to 0. Activations at the output units were recorded. The results presented below report averages over all five networks.

Results

The networks learn many of the simple co-occurrence regularities observed in the data, as well as higher-order co-occurrences. When tested on the object *it*, the most activated verbs are *get*, *hold*, *take* and *have*, which are among the most common verbs with *it* in PCDS (see Table 1). Similarly, *tell*, *make* and *say* are the most activated verbs when networks are tested with the *clause* unit activated in the object position, and they are also among the verbs most commonly associated with a *clause* in the input (Table 2).

However, the networks do not merely learn the relative frequencies of pronouns with verbs. For example, the verbs most activated by the subject *you* are *have* and *get*, neither of which appears in Table 3. The reason for this, we believe, is that the subject *you* is strongly associated with the object *it*, and the object *it*, as mentioned in the previous paragraph, is strongly associated with the verbs *have* and *get*. The difference may be observed most clearly when the network is prompted simultaneously with *you* as the subject and *clause* as the object. In that case, the verb *want* is strongly preferred and, though *get* still takes second place, *tell* and *know* rank third and fourth, respectively — consistent with the results in Table 1. This demonstrates that the network model is sensitive to high-order correlations among words in the input, not merely the first-order correlations between pronoun and verb occurrences.

Conclusion

We have shown that there are statistical regularities in co-occurrences between pronouns and verbs in the speech that children hear from their parents. We have also shown that a simple statistical learner can exploit these regularities, including subtle higher-order regularities that are not obvious in a casual glance at the input data. Furthermore, it can use these regularities to predict the verb in an incomplete sentence.

There are several ways in which this might help children learn verbs. First, pronouns may 'highlight' verbs by consistently bracketing the verb with simple, frequent markers, making it easier to segment the verb from the speech stream. Second, *which* pronouns are used may indicate the animacy, gender, and number of the participants in the action or event that an utterance describes, and the *order* of the pronouns may further indicate temporal sequence or causal direction. Finally, one set of verb-pronoun co-occurrences may lead to another, in the following sense: once the child has learned at least one verb and its pattern of correlations with pronouns, when she hears another verb being used with the same or a similar pattern of correlations, she may hypothesize that the unknown verb is similar to the known verb in some respects. In a sense, this is what our network does — although it does not learn new verbs by this method, it does identify the broad class of a missing verb purely based on its co-occurrences with pronouns and other high-frequency nouns.

We do not claim that the network that we used in Experiment 2 is, by itself and in its current state, an adequate cognitive model. We do claim, however, that it is a suggestive first step. In part, we used the model to study some statistical properties of the data itself, namely higher-order associations. That was not the only purpose of introducing the model, though — the other purpose was to demonstrate that it is possible for a simple statistical learner to use the regularities in the data to predict the class of a missing verb. We would agree that it is very unlikely that children autoassociate simple word patterns; on the other hand, we would insist that it is highly likely that children somehow keep track of lexical co-occurrence patterns. The autoencoder, while assuredly nowhere close to an exact model of the child's processing of parental speech and subsequent lexical learning, is a first step in that direction, to be refined by subsequent research.

The next step in our research is to show that children actually pick up on these regularities. We could be wrong that the lexical co-occurrences are significant — perhaps what children really attend to are referential co-occurrences. While the words *I* and *you* will generally correspond with the referents *me* and *Mommy* / *Daddy* in PCDS, the referents of other pronouns will be more variable. To the extent that children do attend to the lexical regularities, we predict, they should, at a minimum, use pronouns and verbs together with roughly the same frequencies that they hear in their parents' speech. Measuring frequencies is the subject of research in progress using the coded corpus data from Experiment 1 to compute the correlations in the patterns of co-occurrence between verbs and pronouns in parental speech and children's speech.

A further step would be to show that children not only pick up on the regularities, and exhibit them in production, but also use them to comprehend verbs. That is, children should better comprehend ordinary but relatively infrequent verbs when they are presented in frames that are consistent with regularities in the inputs, as opposed to when they are presented in frames that are inconsistent with those regularities. This is a strong test of the network model — to the extent that the model is at all suitable, children should be able to perform at least as well as it does at guessing known, but missing or garbled verbs. In particular, they should be sensitive to higher-order correlations, as our model predicts.

The third and most stringent test is whether children not only exhibit these regularities in production and use them to comprehend known verbs, but also use them to learn novel verbs. Thus, future experiments are planned to measure children's sensitivity to higher-order correlations by testing their preferences for different scenes when primed with sentences containing nonce verbs and pronouns, such as "You blick it" and "I moop him." If children are better able to learn novel verbs when they are presented in pronoun frames consistent with the regularities we have observed, this provides strong support for our hypothesis that pronoun-verb co-occurrences help children learn verbs.

Work is also underway to collect comparable crosslinguistic data from both Japanese and Tamil, relatively verb-heavy languages with frequent argument dropping and case-marked pronouns referring to various levels of social status. Finally, we intend to refine our model along the way, to incrementally approach a fuller and more accurate account of the child's learning processes.

Acknowledgements

This research was supported by NIMH grant number ROI MH 60200. Additional thanks go to our coders, to members of the Cognitive Development Laboratory at IU for useful discussions of these results, and to several anonymous reviewers for helpful comments on an earlier draft.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Cameron-Faulkner, T., Lieven, E. V. M., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27,* 843-873.

Chafe, W. L. (1994). *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Childers, J. B., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology, 37*(6), 739-748.

DuBois, J. W., Kumpf, L. E., & Ashby, W. J. (Eds.). (2003). *Preferred Argument Structure: Grammar as an Architecture for Function.* Amsterdam: John Benjamins Publishers.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed. Vol. 2: The Database). Mahwah, NJ: Lawrence Erlbaum Associates.

Quine, W. V. O. (1980). On what there is. In W. V. O. Quine (Ed.), *From a Logical Point of View* (2nd ed.). Cambridge, MA: Harvard University Press.

Riedmiller, M., & Braun, H. (1993). *A direct adaptive method for faster backpropagation learning: The Rprop algorithm.* Paper presented at the IEEE International Conference on Neural Networks 1993 (ICNN 93), San Francisco, CA.

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40*(1-2), 21-81.

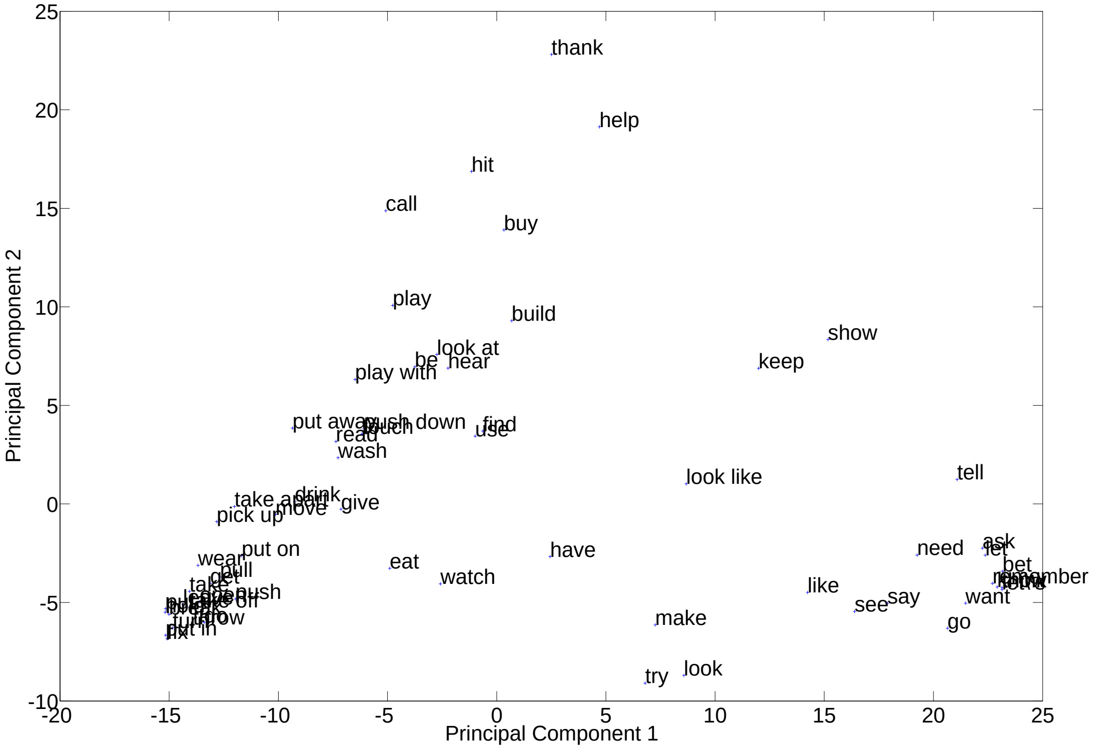

Figure 4: Verbs plotted in the first two principal components of syntactic object space.

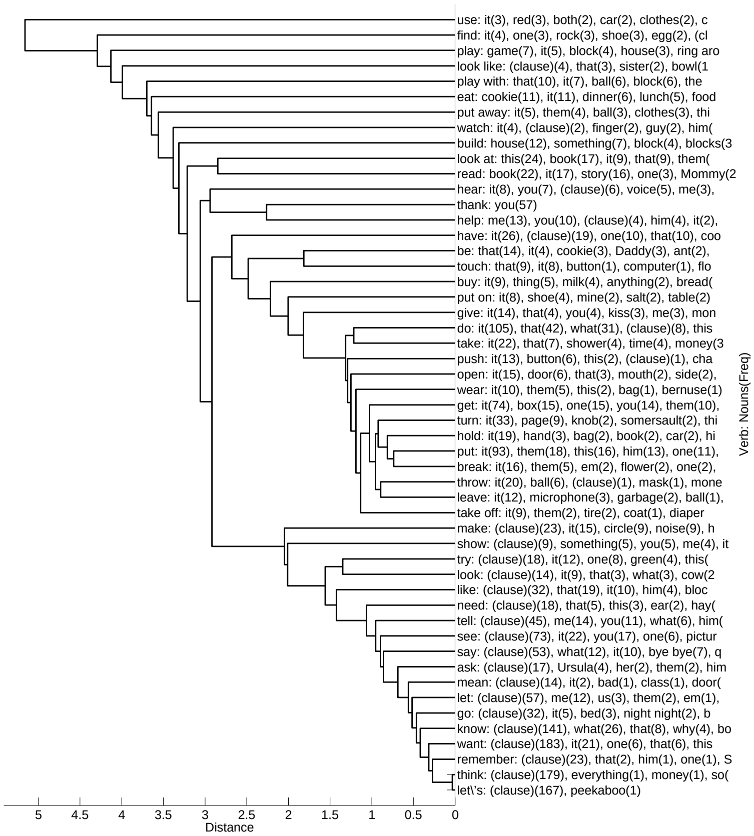

Figure 5: Cluster diagram of verbs in syntactic object space.