

Distributional Clustering of English Words

Fernando Pereira
AT&T Bell Laboratories

Naftali Tishby
Hebrew University

Lillian Lee
Cornell University

April 25, 1993

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Words are represented by the relative frequency distributions of contexts in which they appear, and relative entropy is used to measure the dissimilarity of those distributions. Clusters are represented by “typical” context distributions averaged from the given words according to their probabilities of cluster membership, and in many cases can be thought of as encoding coarse sense distinctions. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical “soft” clustering of the data. Clusters are used as the basis for class models of word cooccurrence, and the models evaluated with respect to held-out test data.

1 Motivation

Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in building statistical language models, particularly in models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example of frequencies of pairs of a transitive main verb and the head noun of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle [Hin90] proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of “similar” events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle’s proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden *senses classes* and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes [Res92], in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or *clusters* c with corresponding cluster membership probabilities $p(c|w)$ for each word w . Most other class-based modeling techniques for natural language rely

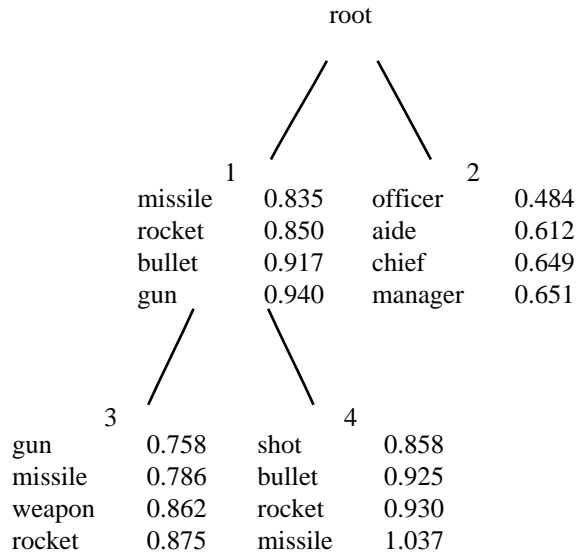


Figure 1: Direct object clusters for *fire*

instead on “hard” Boolean classes [BDPd⁺90]. Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above. Our approach avoids both problems.

2 Problem Setting

In what follows, we will consider two major word classes, \mathcal{V} and \mathcal{N} (for the verbs and nouns in our experiments) and a single relation between them (the main verb-head of direct object relation in our experiments) sampled by the frequencies f_{vn} of occurrence of particular pairs (v, n) in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle’s parser Fidditch [Hin93]. More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger [Chu88] and of tools for regular expression pattern matching on tagged corpora [Yar92]. We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like “say”.

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n -ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between those hidden units.

For the noun classification problem, the empirical distribution of a noun n can then be given by the conditional density $p_n(v) = f_{vn} / \sum_v f_{vn}$. The problem we study is how to use the p_n to classify the $n \in \mathcal{N}$. Our classification method will construct a set \mathcal{C} of clusters and cluster membership probabilities $p(c|n)$. Each cluster c is associated to a cluster *centroid* p_c , which is discrete density over \mathcal{V} obtained by averaging appropriately the p_n .

3 Measuring Distributional Similarity

We work here with a measure of distributional *dissimilarity* rather than similarity. The dissimilarity between two nouns n and n' will be simply the relative entropy (Kullback-Leibler distance) of the corresponding conditional verb distributions

$$D(p_n \parallel p_{n'}) = \sum_v p_n(v) \ln \frac{p_n(v)}{p_{n'}(v)}$$

This is a well known measure of dissimilarity between densities, which is zero just in case the densities are identical and increases as the likelihood of the first density being an empirical sample drawn according to the second density decreases. In information-theoretic terms, $D(f \parallel f')$ measures how inefficient on average it would be to use a code based on f' to encode a variable distributed according to f . With respect to our problem, $D(p_n \parallel p_c)$ thus gives us the loss of information in using cluster centroid p_c instead of the actual distribution for word p_n when modeling the distributional properties of n . We will also show that relative entropy is a natural measure of dissimilarity between distributions for clustering because its minimization leads to cluster centroids that are a simple weighted average of member distributions.

One technical difficulty is that $D(f \parallel f')$ is not defined (infinite) when $f'(x) = 0$ but $f(x) > 0$. We could sidestep this problem (as we did initially) by smoothing zero frequencies appropriately [CG91]. However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes. It turns out that the problem is avoided by our clustering technique, which never needs to compute the dissimilarity between individual word distributions, but only between a word distribution and average distributions (cluster centroids) that are guaranteed to be nonzero whenever the word distributions are (except in numeric underflow situations). This is an important advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.

4 THEORETICAL BASIS

In its most general form, the problem we are considering here can be stated as follows. There is a large set of linguistic objects, typically words, which whose occurrence in certain textual contexts (for instance, sentences, grammatical constructions, n -grams) has been sampled. Our goal is to organize the objects using only the information provided by the samples.

For the objectives of the present paper, however, we will only consider the more particular question outlined in Section 2. The problem can then be seen as the learning of a joint distribution of pairs from a large training set of independently drawn pairs. The pair coordinates come from two large sets of objects \mathcal{X} and \mathcal{Y} , with no preexisting topological or metric structure, and the training data is a sequence S of N independently drawn pairs

$$S_i = (x_i, y_i) \quad 1 \leq i \leq N .$$

From a learning perspective, this problem falls somewhere in between unsupervised and supervised learning. As in unsupervised learning, the goal is to learn the underlying distribution of the data. In contrast to most unsupervised learning settings, however, the objects involved have no internal structure or attributes allowing them to be compared with each other. Instead, the only information about the objects is the statistics of their joint appearance. These statistics can thus be seen as a weak form of object labelling analogous to supervision. As in other learning problems, the proof of the utility of the model is in its ability to generalize to data beyond the training examples.

As outlined in Section 2, objects will be characterized by conditional empirical distributions. Each pair (x, y) with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ occurs with a certain (possibly zero) frequency f_{xy} in S . Each object x will then be characterized by the conditional distribution

$$p_x(y) \stackrel{\text{def}}{=} p(y|x) = \frac{f_{xy}}{\sum_{y \in \mathcal{Y}} f_{xy}} .$$

This distribution can be thought of as the “statistical signature” of x in terms of its occurrence with elements of \mathcal{Y} in the chosen contexts. Similarly one can construct the signatures of the objects y :

$$p_y(x) \stackrel{\text{def}}{=} p(x|y) = \frac{f_{xy}}{\sum_{x \in \mathcal{X}} f_{xy}}$$

so that to each member of \mathcal{X} corresponds a distribution on \mathcal{Y} , and to each member of \mathcal{Y} , a distribution on \mathcal{X} .

4.1 Distributional Clustering

Distributions, unlike abstract objects, can be combined and averaged. This enables us to cluster objects through their associated distributions rather than having to rely on specific structural or metric properties of the objects.

For a given set of clusters \mathcal{C} we would like to find cluster distribution centroids, $p(y|c)$, such that each object distribution can be approximately decomposed as

$$\hat{p}_x(y) = \hat{p}(y|x) = \sum_{c \in \mathcal{C}} p(c|x)p(y|c) \quad .$$

Such a decomposition can be written in a more symmetric form as

$$\begin{aligned} \hat{p}(x, y) &= \sum_{c \in \mathcal{C}} p(c, x)p(y|c) \\ &= \sum_{c \in \mathcal{C}} p(c)p(x|c)p(y|c) \end{aligned} \quad (1)$$

which is taken as our basic clustering model.

To achieve this decomposition we need to answer two questions involving the combination of two complementary variational principles. The first question concerns the functional forms of the cluster membership distributions $p(c|x)$ and the centroid distributions $p(y|c)$, and the second concerns the goodness of fit of the model to the observed data. Clearly, these two problems are connected.

Goodness of fit is determined by the likelihood of the observations given the model. The maximum likelihood (ML) estimation principle is thus the natural tool to determine the centroid distributions $p(y|c)$.

As for the membership probabilities, they must be determined solely by the relevant measure of object-to-cluster similarity, which in the present work is the relative entropy between the object and cluster centroid distributions. Since no other information is available, the membership is determined by maximizing the configuration entropy subject to the average distortion as a constraint. We shall show that, with the maximum entropy (ME) membership distribution, ML estimation is equivalent to the minimization of the average distortion of the data. The combined entropy maximization and distortion minimization is carried out by a two-stage iterative process similar to the standard EM method [DLR77]. The first stage of an iteration is a maximum likelihood, or minimum distortion, estimation of the cluster centroids given fixed membership probabilities. In the second iteration stage, the entropy of the membership distribution is maximized with a fixed average distortion. This joint optimization searches for a *saddle point* in the distortion-entropy parameters, which is equivalent to the minimization of a linear combination of the two known as *free energy* in statistical mechanics. This analogy with statistical mechanics is not coincidental and provide us with a better understanding of the clustering procedure itself.

4.1.1 Maximum Likelihood Cluster Centroids

For the maximum likelihood argument, we start by estimating the likelihood of a sequence S of N independent observations, each a pair (x_i, y_i) of objects cooccurring in some context. Using (1), the sequence’s model

likelihood can be written as

$$\mathcal{L}(S) = \hat{p}(S) = \prod_{i=1}^N \sum_{c \in \mathcal{C}} p(c) p(x_i|c) p(y_i|c) \quad .$$

Their *log-likelihood* is thus

$$l(S) = \log \hat{p}(S) = \sum_{i=1}^N \log \sum_{c \in \mathcal{C}} p(c) p(x_i|c) p(y_i|c) \quad .$$

Fixing the number of clusters (model size) $|\mathcal{C}|$, we seek to maximize the likelihood with respect to the centroid distributions $p(x|c)$ and $p(y|c)$. The variation of the log-likelihood with respect to these distributions is

$$\delta l(S) = \sum_{i=1}^N \frac{1}{\hat{p}(x_i, y_i)} \sum_{c \in \mathcal{C}} p(c) \begin{pmatrix} p(y_i|c) \delta p(x_i|c) \\ + \\ p(x_i|c) \delta p(y_i|c) \end{pmatrix} \quad (2)$$

where the normalization of $p(x|c)$ and $p(y|c)$ is assumed preserved. Using Bayes's formula we can simplify this expression,¹ since

$$p(x_i|c) p(y_i|c) = \frac{p(c|x_i, y_i)}{p(c)} \hat{p}(x_i, y_i) \quad ,$$

or

$$\frac{1}{\hat{p}(x_i, y_i)} = \frac{p(c|x_i, y_i)}{p(c) p(x_i|c) p(y_i|c)} \quad (3)$$

for any c . Substituting (3) into (2) we obtain

$$\delta l(S) = \sum_{i=1}^N \sum_{c \in \mathcal{C}} p(c|x_i, y_i) \begin{pmatrix} \delta \log p(x_i|c) \\ + \\ \delta \log p(y_i|c) \end{pmatrix} \quad (4)$$

using $\delta \log p = \delta p/p$. This expression is particularly useful when the cluster distributions, $p(x|c)$ and $p(y|c)$ are of an exponential form, precisely what is provided by the maximum entropy step.

At this point, however, it is necessary to specify the clustering model more carefully. The simplest possibility, and the one implemented in this paper, is to make two *independent* sets of clusters, one in the \mathcal{X} space, denoted \mathcal{C}_x , and another for the \mathcal{Y} space. This scheme we call the *asymmetric model*. In the asymmetric model the cluster membership is determined by only one component (for example, x) where the clusters correspond to a distribution on the other component ($p(y|c)$). This scheme simplifies the estimation significantly by dealing with a single component. The main disadvantage of this model is that the joint likelihood, $p(x, y)$, has two different expressions, which are not necessarily consistent, one through the \mathcal{X} clusters, and another through \mathcal{Y} .

An alternative scheme, which avoids the likelihood consistency problem, is the *symmetric model*

$$p(x, y) = \sum_{c_x} \sum_{c_y} p(c_x, c_y) p(x|c_x) p(y|c_y) \quad .$$

This model requires the estimation of the joint cluster probabilities, $p(c_x, c_y)$, and will be discussed elsewhere. Here we proceed with the asymmetric model.

¹As usual in clustering models [DH73], we assume that the model distribution and the empirical distribution are interchangeable at the solution of the parameter estimation equations, since the model is assumed to be able to represent correctly the data at that solution point. In practice, the data may not come exactly from the chosen model class, but the model obtained by solving the estimation equations may still be the closest one to the data.

4.1.2 Maximum Entropy Cluster Membership

In the asymmetric model the variations $\delta p(x|c)$ and $\delta p(y|c)$ are not independent. The centroid distributions should then be determined via ML so that the membership probabilities $p(c|x)$ minimize some average distortion between $p(y|c)$ and $p(y|x)$. It is therefore enough to carry out this minimization first, which deals with the variation of $p(y|c)$ as well.

Generally, given any dissimilarity measure $D[x|c]$ between object and centroid distributions p_x and p_c , the average cluster distortion is

$$\langle D \rangle = \sum_x \sum_c p(c|x) D[x|c] \quad .$$

If we maximize the cluster entropy

$$H = - \sum_x \sum_c p(c|x) \log p(c|x) \quad ,$$

subject to normalization and average distortion constraints, we obtain the following standard exponential expressions for the class and membership distributions

$$p(x|c) = \frac{1}{Z_c} \exp(-\beta D[x|c]) \quad (5)$$

$$p(c|x) = \frac{1}{Z_x} \exp(-\beta D[x|c]) \quad (6)$$

where the normalization sums (partition functions) are $Z_c = \sum_x \exp(-\beta D[x|c])$ and $Z_x = \sum_c \exp(-\beta D[x|c])$. Notice that $D[x|c]$ need not be symmetric for this derivation, and the two distributions are simply related by the Bayes formula.

Returning to (4), substituting in the exponential form (5) for $p(x|c)$, and taking into account the assumption for the asymmetric model that the cluster membership probabilities are independent of the \mathcal{Y} space, we obtain

$$\delta l(S) = - \sum_{i=1}^N \sum_{c \in \mathcal{C}} p(c|x_i) \delta \beta D[x_i|c] + \delta \log Z_c \quad (7)$$

where the variation of $p(y|c)$ is now included in the variation of $D[x|c]$ and thus should not be considered independently.

For a large enough sample, we may replace the sum over observations in (7) by the average over \mathcal{X}

$$\delta l(S) = - \sum_x p(x) \sum_{c \in \mathcal{C}} p(c|x) \delta \beta D[x|c] + \delta \log Z_c$$

which, applying Bayes's rule, becomes

$$\delta l(S) = - \sum_{c \in \mathcal{C}} \frac{1}{p(c)} \sum_x p(x|c) \delta \beta D[x|c] + \delta \log Z_c \quad (8)$$

At the log-likelihood maximum, the variation (8) is required to vanish. We will see below that our chosen dissimilarity measure has the desirable property that $\delta \log Z_c$ vanishes at the likelihood maximum as well. Thus the log-likelihood can be maximized by minimizing the average distortion with respect to the class centroids while keeping the class membership fixed

$$\sum_c \frac{1}{p(c)} \sum_x p(x|c) \delta D[x|c] = 0 \quad ,$$

or, sufficiently, if each of the inner sums vanish

$$\sum_c \sum_x p(x|c) \delta D[x_i|c] = 0 \quad (9)$$

Furthermore, it is a natural requirement that the cluster centroid distributions should be weighted averages of object distributions

$$p(y|c) = \sum_x p(x|c)p(y|x) \quad (10)$$

We shall see below that those two conditions are satisfied by our chosen dissimilarity measure, the relative entropy

$$D[x|c] = D(p_x \parallel p_c) = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|c)} \quad (11)$$

In fact, it is possible to show, and we will do so elsewhere, that the relative entropy is the *only* such function, up to mild regularity conditions.

4.1.3 Minimizing the Average KL Distortion

We first show that the minimization of the relative entropy yields the natural expression for the centroids (10). To minimize the average distortion (9), we observe that the variation of (11) with respect to the centroid distributions $p(x|c)$, with each centroid distribution normalized by the Lagrange multiplier λ_c , is given by

$$\begin{aligned} \delta D[x|c] &= \delta \left(\begin{array}{c} -\sum_y p(y|x) \log p(y|c) \\ + \\ \lambda_c (\sum_y p(y|c) - 1) \end{array} \right) \\ &= \sum_y \left(-\frac{p(y|x)}{p(y|c)} + \lambda_c \right) \delta p(y|c) \quad . \end{aligned}$$

Substituting this expression into (9), we obtain

$$\sum_c \sum_x \sum_y \left(-\frac{p(y|x)p(x|c)}{p(y|c)} + \lambda_c \right) \delta p(y|c) = 0 \quad ,$$

or as the final reestimation formula, using the fact that the $\delta p(x|c)$ are now independent for each cluster c :

$$p(y|c) = \sum_x p(x|c)p(y|x) \quad .$$

This last form for the cluster centroid distributions is simply a weighted average of the object distributions, as required.

It remains to show that the variation $\delta \log Z_c$ vanishes on that solution. Indeed,

$$\begin{aligned} \delta \log Z_c &= -\frac{\beta}{Z_c} \sum_x e^{-\beta D[x|c]} \delta D[x|c] \\ &= -\beta \sum_x p(x|c) \delta D[x|c] \quad , \end{aligned}$$

which vanishes precisely with the variation of the mean distortion (9).

4.1.4 The Free Energy Function

The combined minimum distortion and maximum entropy optimization is equivalent to the minimization of a single function, the *free energy* F . From the expression for the association entropy

$$H = - \sum_x \sum_c p(c|x) \log p(c|x) = \beta \langle D \rangle - \beta F \quad ,$$

where

$$F = -\beta^{-1} \sum_x \log \sum_c \exp(-\beta D[x|c]) \quad . \quad (12)$$

The free energy determines both the distortion and the membership entropy through

$$\langle D \rangle = \frac{\partial \beta F}{\partial \beta} \quad (13)$$

$$S = -\frac{\partial F}{\partial T} \quad , \quad (14)$$

where the *temperature* $T = \beta^{-1}$.

The most important property of the free energy is that its minimum determines the balance between the “disordering” maximum entropy and “ordering” distortion minimization in which the system is most likely to be found. In fact the probability to find the system at a given configuration is exponential in F

$$P \propto \exp -\beta F \quad , \quad (15)$$

so a system is most likely to be found in its minimal free energy configuration.

The analogy with statistical mechanics suggest a *deterministic annealing* procedure for the clustering [RGF90], in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter β (decreasing the temperature T) as we explain in the next section.

4.2 Hierarchical Clustering

The statistics of natural languages is inherently ill defined. Because of Zipf’s law, there is never enough data for a reasonable estimation of joint object distributions. For any corpus size, however, there is a level of object discrimination supported by the data, which also depends on the specificity of word usage in the corpus. Clearly, if fine distinctions between objects are not used, no amount of data with the same statistical properties can supply the missing distinctions. The appropriate number of clusters for a certain corpus is thus determined by the combination of the amount of data and of the specificity in usage in the corpus, which can be thought as inversely related to a level of “noise”. It is thus natural to consider a hierarchical clustering scheme to construct clusters at the appropriate level of detail.

Our formulation so far has provided all the ingredients for such a scheme. Consider the *free energy* for a given number of clusters, given by (12) as function of the temperature parameter $T = \beta^{-1}$. As $\beta \rightarrow \infty$ ($T \rightarrow 0$) the membership distribution (6) becomes increasingly sharper (the clusters become “harder”). Each object becomes associated just with its nearest centroid, and the membership entropy approaches zero. In this *hard clustering limit*, the free energy, and therefore the distortion no longer changes with decreasing T , as a result of (13). The clusters are therefore effectively *frozen* below a certain temperature.

In a many-cluster system at very small β , membership entropy will be high, since there are more possible associations. At this *high temperature* limit the free energy is mostly entropy and thus the free energy is higher for more clusters. At the low T limit, on the other hand, the entropy is negligible compared to the distortion, and since there are more clusters among which to divide the objects, the average distortion is lower when more clusters are allowed. Since both free energies are continuous monotonically decreasing functions of the temperature, there will be a unique *critical* point in between where the two free energies have equal value.

Since the value of the free energy determines the configuration through (15), the typical configuration corresponds to the the minimal free energy. If the number of clusters is allowed to vary, the system is likely to move spontaneously from one free energy curve to another with higher number of clusters at the critical point where they intersect. Such a process is known as a *phase transition* in statistical physics, and reflects a discontinuous change in the system’s most likely configuration. To observe these transition for clustering it is necessary to follow the free energy of two solutions of consecutive number of clusters as functions of the decreasing temperature. An effective way of doing this is through *deterministic annealing*, as proposed by Rose *et al.* [RGF90]. In our implementation we have adopted this scheme, where in addition to the empirical calculation of the free energy as function of the temperature, we determine the actual phase transitions by recognizing at what temperature replicated cluster centroids become distinct.

5 Clustering Examples

All our experiments involve the asymmetric model described in the previous section. As explained there, our clustering procedure yields for each value of β a set C_β of cluster centroids that locally minimize the free energy F , and the asymmetric model for β estimates the conditional verb distribution for a noun n by

$$\hat{p}_n = \sum_{c \in C_\beta} p(c|n)p_c$$

where $p(c|n)$ also depends on β .

As a first experiment, we used our method to classify the 64 nouns appearing most frequently as heads of direct objects of the verb “fire” in one year (1988) of Associated Press newswire.² In this corpus, the chosen nouns appear as direct object heads of a total of 2147 distinct verbs. Thus, each noun is represented by a density over the 2147 verbs.

Figure 1 shows the five words most similar to each cluster centroid for the four clusters resulting from the first two cluster splits. It can be seen that first split separates the objects corresponding to the weaponry sense of “fire” (cluster 1) from the ones corresponding to the personnel action (cluster 2). The second split then further refines the weaponry sense into a projectile sense (cluster 3) and a gun sense (cluster 4), although that split is somewhat less sharp, possibly because not enough distinguishing contexts occur in the corpus.

Figure 2 shows the four closest nouns to the centroid of each of a set of hierarchical clusters derived from verb-object pairs involving the 1000 most frequent nouns in the June 1991 electronic version of Grolier’s Encyclopedia (10 million words). This corpus had previously been automatically tagged with parts of speech [Chu88]. With the help of concordancing and pattern-matching tools [Yar92], we extracted from the corpus a set of verb-noun pairs likely to correspond to the main verb and to the head of the direct object of the verb. Since the problem of identifying with certainty actual direct object noun phrases in general text is beyond the abilities of regular pattern matching (or even of existing parsers), we used a variety of additional information, including verb subcategorization information in dictionaries and random samples of candidate direct objects to refine our pattern matching rules and filter the results to remove unlikely verb-direct object configurations and also likely duplicates resulting from sentence and article duplication in the initial corpus. We also addressed the problem that in noun phrases like “a type of building” the syntactic head “type” is semantically uninformative, so we tabulate instead the “semantic head” “building.”

While we have not yet developed sufficiently detailed methods for evaluating such cluster sets, it is interesting to see how major semantic classes (locations, times, actions, physical structures, animated beings, scalar variables) emerge from the procedure. Some of the less felicitous cluster associations result from data collection artifacts that our pattern matching techniques failed to identify. For example, the occurrence of “number” close to many of the cluster centroids comes from failing to recognize the semantic head of “number of” noun phrases, and the occurrence of “year” in some inappropriate clusters comes from misidentifying time adverbials like “the previous year” as verb direct objects.

²The verb-object pairs for this example were collected by Don Hindle using his parser Fidditch, and this particular subset selected by Mats Rooth.

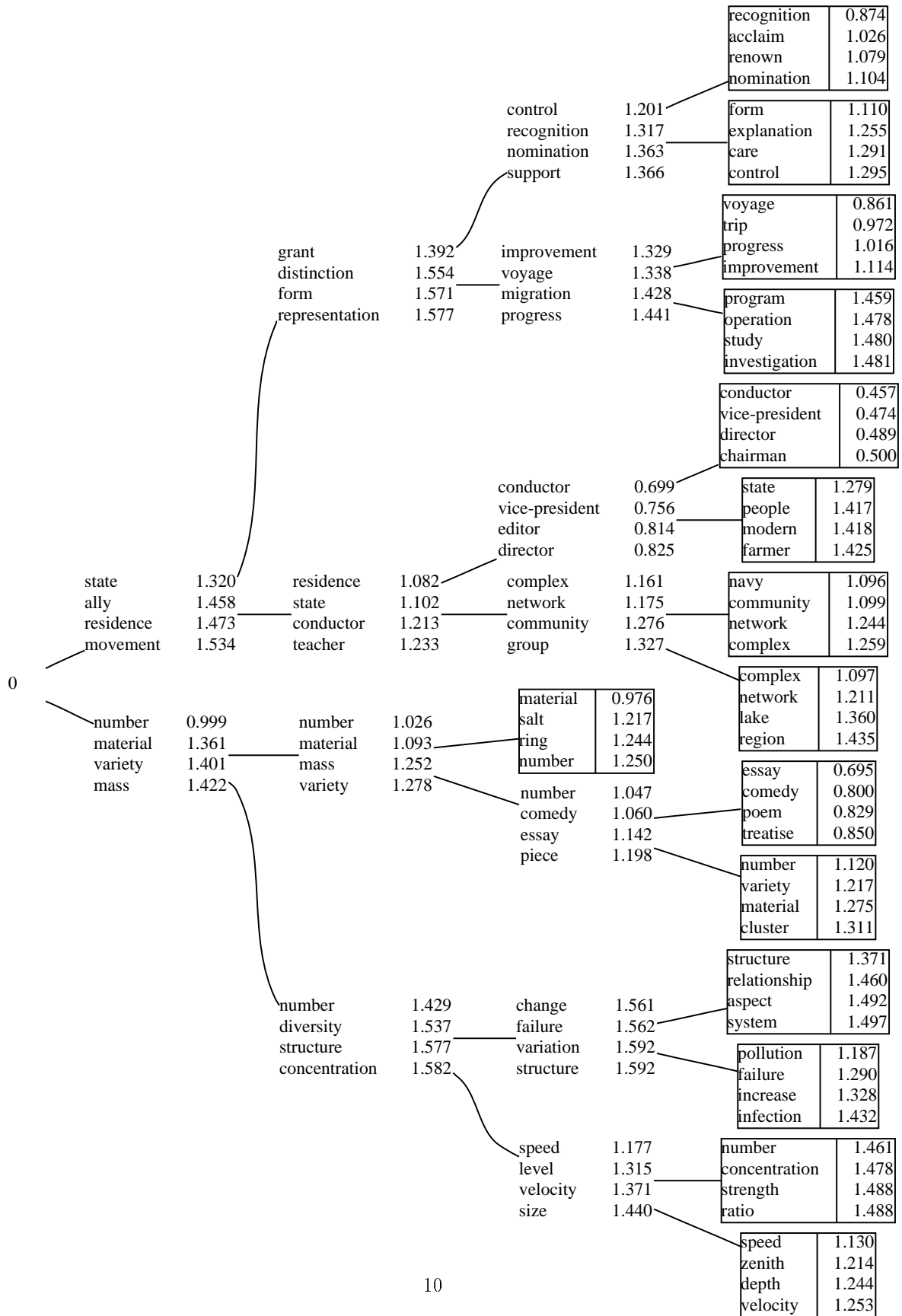


Figure 2: Noun Clusters for Grolier's Encyclopedia

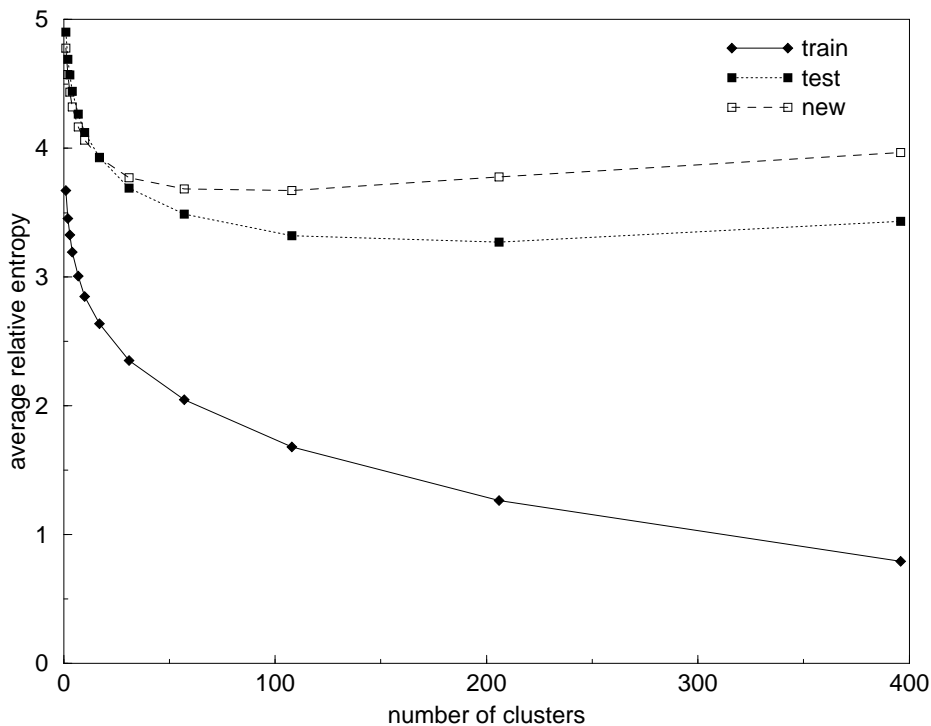


Figure 4: Asymmetric Model Evaluation, AP88 Verb-Direct Object Pairs

6 Model Evaluation

The preceding qualitative discussion provides some indication of what aspects of distributional relationships may be discovered by clustering. However, we also need to evaluate clustering more rigorously as a basis for models of distributional relationships. So, far, we have looked at two kinds of measurements of model quality: (i) relative entropy between held-out data and the asymmetric model, and (ii) performance on the task of deciding which of two verbs is more likely to take a given noun as direct object when the data relating one of the verbs to the noun has been deleted from the data used to build the model.

The evaluation described below was performed on the largest data set we have worked with so far, extracted from 44 million words of 1988 Associated Press newswire with the pattern matching techniques discussed in Section 5. This collection process yielded 1112041 verb-object pairs. We selected then the subset involving the 1000 most frequent nouns in the corpus for clustering, and randomly divided it into a training set of 756721 pairs and a test set of 81240 pairs. Figure 3 shows the closest nouns to the cluster centroids after a few subdivisions while training on that data set.

6.1 Relative Entropy

Figure 4 plots the average relative entropy of several data sets to asymmetric clustered models of different sizes, given by

$$\sum_n D(t_n || \hat{p}_n)$$

where t_n is the relative frequency distribution of verbs taking n as direct object in the test set.³

³We do not weight the terms of the sum by the relative frequencies of nouns because we are interested in the accuracy of distributional modeling independently of the frequencies of particular nouns.

The annealing procedure was used to cluster the conditional verb distributions p_n of the training set described above. At each critical value of β , we computed the relative entropy with respect to the asymmetric model based on C_β of the training set (set *train*), of the randomly selected held-out test set (set *test*), and of held-out data for a further 1000 nouns that were not clustered (set *new*). Figure 4 plots those relative entropy values against the number of clusters in the model. Unsurprisingly, the training set relative entropy decreases monotonically. The test set relative entropy decreases to a minimum at 206 clusters, and then starts increasing, suggesting that larger models are overtrained.

The new noun test set is intended to test whether clusters based on the 1000 most frequent nouns are useful classifiers for the selectional properties of nouns in general. Since the nouns in the test set pairs do not occur in the training set, we do not have their cluster membership probabilities that are needed in the asymmetric model. Instead, for each noun n in the test set, we classify it with respect to the clusters by setting

$$p(c|n) = \exp -\beta D(p_n||c)/Z_n$$

where p_n is the empirical conditional verb distribution for n given by the test set. These cluster membership estimates were then used in the asymmetric model and the test set relative entropy calculated as before. As the figure shows, the cluster model provides over one bit of information about the selectional properties of the new nouns, but the overtraining effect is even sharper than for the held-out data involving the 1000 clustered nouns.

6.2 Decision Task

We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to disambiguation in language analysis. The task consists judging which of two verbs v and v' is more likely to take a given noun n as object, when all occurrences of (v, n) in the training set were deliberately deleted. Thus this test evaluates how well the models reconstruct missing data in the verb distribution for n from the cluster centroids close to n .

The data for this test was built from the training data for the previous one in the following way, based on a suggestion by Dagan *et al.* [DMM92]. A small number (104) of (v, n) pairs with a fairly frequent verb (between 500 and 5000 occurrences) was randomly picked, and all occurrences of each pair in the training set were deleted. The resulting training set was used to build a sequence of cluster models as before. Each model was used to decide which of two verbs v and v' are more likely to appear with a noun n where the (v, n) data was deleted from the training set, and the decisions compared with the corresponding ones derived from the original event frequencies in the initial data set. More specifically, for each deleted pair (v, n) and each verb v' that occurred with n in the initial data either at least twice as frequently or at most half as frequently as v , we compared the sign of $\log \hat{p}_n(v)/\hat{p}_n(v')$ with that of $\log p_n(v)/p_n(v')$ for the initial data set. The error rate for each model is simply the proportion of sign disagreements in the selected (v, n, v') triples. Figure 5 shows the error rates for each model for all the selected (v, n, v') (*all*) and for just those *exceptional* triples in which the log frequency ratio of (n, v) and (n, v') differs from the log marginal frequency ratio of v and v' . In other words, the exceptional cases are those in which predictions based just on the marginal frequencies, which the initial one-cluster model represents, would be consistently wrong.

Here too we see some overtraining for the largest models considered, although not for the exceptional verbs.

7 Conclusions and Further Work

We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words. The resulting clusters are intuitively informative, and can be used to construct class-based word cooccurrence models with substantial predictive power.

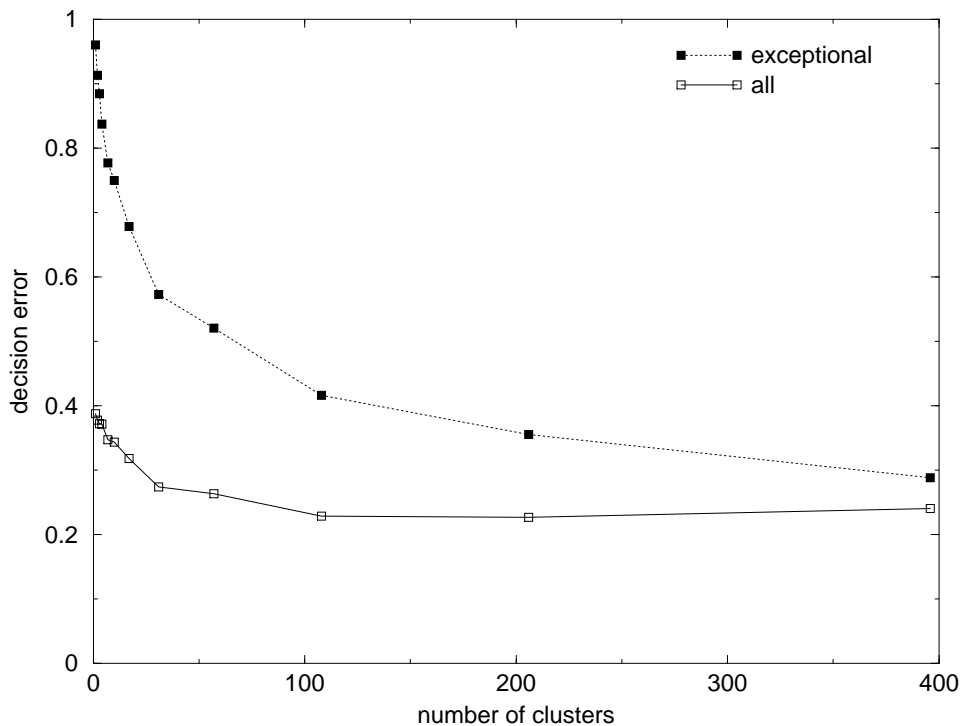


Figure 5: Pairwise Verb Comparisons, AP88 Verb-Direct Object Pairs

While the clusters derived by the proposed method seem in many cases semantically significant, this intuition needs to be grounded in a more rigorous assessment. In addition to predictive power evaluations of the kind we have already carried out, it might be worth comparing automatically-derived clusters with human judgements in a suitable experimental setting.

Moving further in the direction of class-based language models, we plan to consider additional distributional relations (for instance, adjective-noun) and apply the results of clustering to the grouping of lexical associations in lexicalized grammar frameworks such as stochastic lexicalized tree-adjoining grammars [Sch92].

8 Acknowledgments

We would like to thank Don Hindle for making available the 1988 Associated Press verb-object data set, the Fidditch parser and a verb-object structure filter, Mats Rooth for selecting the objects of “fire” data set and many discussions, David Yarowsky for help with his stemming and concordancing tools, and Ido Dagan for suggesting ways of testing cluster models.

References

[BDPd+90] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, pages 283–298, Paris, France, March 1990.

- [CG91] Kenneth W. Church and William A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54, 1991.
- [Chu88] Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas, 1988. Association for Computational Linguistics, Morristown, New Jersey.
- [DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, New York, 1973.
- [DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
- [DMM92] Ido Dagan, Shaul Markus, and Shaul Markovitch. Contextual word similarity and the estimation of sparse lexical relations. Submitted for publication, 1992.
- [Hin90] Donald Hindle. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, Pennsylvania, 1990. Association for Computational Linguistics, Morristown, New Jersey.
- [Hin93] Donald Hindle. A parser for text corpora. In B.T.S. Atkins and A. Zampoli, editors, *Computational Approaches to the Lexicon*. Oxford University Press, Oxford, England, 1993. To appear.
- [Res92] Philip Resnik. WordNet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-Based Natural-Language-Processing Techniques*, San Jose, California, July 1992.
- [RGF90] Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.
- [Sch92] Yves Schabes. Stochastic lexicalized tree-adjointing grammars. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992.
- [Yar92] David Yarowsky. CONC: Tools for text corpora. Technical Memorandum 11222-921222-29, AT&T Bell Laboratories, 1992.

A Clusters from Grolier’s Encyclopedia

The listing below shows in some detail the 28 cluster solution for Grolier’s 1000 most frequent nouns. The leaf clusters in Figure 2 are basically the parents of those more fine-grained clusters. For each cluster, it shows the value of β at which the new cluster was identified, the cluster size $\sum_x p(x \in c)$, the five verbs with highest conditional probability in the cluster centroid distribution, the ten nouns with lowest KL distance to the cluster centroid, and the ten nouns with highest membership probability in the cluster. It is important to keep in mind that high cluster membership probability for a word does not necessarily mean that the word is close to the cluster’s centroid, but only that it is *closer* to that centroid than to others. Furthermore, the exponential form of membership probabilities sharpens the contrast between clusters for high β . Thus, a word may appear below as having close to 1 membership probability in a particular cluster, and yet be relatively far in terms of KL distance from the cluster’s centroid.

Cluster 19
size = 70.0808
 $\beta = 4.233224$

most likely verbs	
produce	0.054
include	0.044
bear	0.038
contain	0.034
make	0.025

closest nouns		members	
number	1.120	resemblance	0.994
variety	1.217	portrait	0.984
material	1.275	flower	0.976
cluster	1.311	scene	0.967
set	1.353	cluster	0.957
combination	1.402	landscape	0.957
modern	1.405	pair	0.920
equipment	1.410	fruit	0.918
drawing	1.414	arm	0.916
pair	1.443	son	0.868

Cluster 29
size = 17.88524
 $\beta = 4.253067$

most likely verbs	
pay	0.091
average	0.060
reach	0.057
exceed	0.046
stand	0.042

closest nouns		members	
velocity	1.216	mm	1.000
percent	1.338	tax	1.000
m	1.399	m	1.000
cm	1.447	tribute	1.000
deg	1.474	cm	1.000
cost	1.509	north	1.000
tribute	1.533	debt	1.000
tax	1.570	km	0.999
sum	1.652	sum	0.999
price	1.666	deg	0.999

Cluster 31
size = 39.72022
 $\beta = 4.253067$

most likely verbs	
enter	0.053
dominate	0.029
occupy	0.026
include	0.025
create	0.023

closest nouns		members	
region	1.326	politic	0.985
lake	1.393	area	0.958
half	1.399	market	0.946
area	1.416	council	0.939
legislature	1.475	land	0.938
village	1.477	region	0.934
space	1.488	home	0.922
body	1.497	atmosphere	0.896
valley	1.500	interior	0.895
network	1.502	half	0.890

Cluster 26
size = 56.49065
 $\beta = 4.233224$

most likely verbs	
become	0.473
remain	0.028
include	0.019
establish	0.010
make	0.009

closest nouns		members	
conductor	0.457	minister	1.000
vice-president	0.474	chairman	1.000
director	0.489	professor	0.999
chairman	0.500	director	0.999
professor	0.512	effective	0.999
commander	0.563	president	0.999
center	0.573	leader	0.999
secretary	0.586	center	0.999
editor	0.592	king	0.998
bishop	0.626	secretary	0.998

Cluster 30
size = 23.52949
 $\beta = 4.253067$

most likely verbs	
reach	0.264
commit	0.050
attain	0.030
keep	0.025
set	0.025

closest nouns		members	
zenith	0.862	suicide	1.000
height	0.903	point	0.999
depth	0.964	peak	0.998
maturity	0.991	maturity	0.997
peak	1.017	height	0.997
speed	1.069	crime	0.997
length	1.196	zenith	0.996
sea	1.221	planet	0.993
extent	1.254	depth	0.992
age	1.328	earth	0.990

Cluster 32
size = 38.3321
 $\beta = 4.253067$

most likely verbs	
form	0.143
contain	0.032
include	0.022
become	0.021
cross	0.020

closest nouns		members	
complex	0.953	border	0.988
ring	1.049	barrier	0.985
nucleus	1.077	edge	0.978
core	1.124	mountain	0.971
strip	1.230	continent	0.969
chain	1.258	boundary	0.968
edge	1.280	cavity	0.955
layer	1.291	wall	0.929
network	1.293	bond	0.923
rock	1.326	basis	0.901

Cluster 33
size = 33.38036
 $\beta = 4.253067$

most likely verbs	
begin	0.110
perform	0.039
conduct	0.036
include	0.023
complete	0.019

closest nouns		members	
operation	1.116	experiment	0.999
study	1.157	task	0.998
investigation	1.181	research	0.997
project	1.271	career	0.997
construction	1.302	operation	0.995
experiment	1.323	negotiation	0.995
program	1.369	duty	0.988
negotiation	1.373	publication	0.984
movement	1.420	conquest	0.983
conquest	1.503	construction	0.983

Cluster 35
size = 19.24459
 $\beta = 4.253067$

most likely verbs	
meet	0.066
settle	0.044
face	0.036
create	0.032
pose	0.030

closest nouns		members	
problem	1.109	dispute	1.000
challenge	1.373	threat	1.000
crisis	1.420	question	1.000
controversy	1.502	hostility	1.000
conflict	1.524	controversy	1.000
unemployment	1.562	challenge	0.999
demand	1.563	crisis	0.999
debate	1.565	discrimination	0.999
difficulty	1.593	requirement	0.999
threat	1.675	problem	0.997

Cluster 37
size = 33.16354
 $\beta = 4.253067$

most likely verbs	
produce	0.063
carry	0.052
provide	0.035
lay	0.029
contain	0.028

closest nouns		members	
material	1.154	egg	1.000
oxygen	1.195	foundation	0.999
water	1.224	blood	0.997
impulse	1.269	ball	0.995
heat	1.331	message	0.987
quantity	1.372	signal	0.985
electricity	1.381	impulse	0.984
nutrient	1.382	load	0.977
gas	1.423	datum	0.964
energy	1.427	electricity	0.953

Cluster 34
size = 24.63102
 $\beta = 4.253067$

most likely verbs	
spend	0.066
launch	0.048
take	0.045
pass	0.036
begin	0.019

closest nouns		members	
year	1.434	month	1.000
rest	1.442	day	1.000
day	1.464	year	1.000
measure	1.482	hour	1.000
month	1.527	satellite	0.999
hour	1.553	offensive	0.998
examination	1.621	constitution	0.998
offensive	1.675	bill	0.995
attack	1.710	examination	0.994
resolution	1.724	legislation	0.993

Cluster 36
size = 33.52218
 $\beta = 4.253067$

most likely verbs	
cause	0.125
produce	0.046
reduce	0.036
suffer	0.033
prevent	0.032

closest nouns		members	
pollution	0.996	kg	1.000
increase	1.103	damage	0.999
failure	1.132	defeat	0.999
infection	1.174	pain	0.998
loss	1.184	death	0.998
reduction	1.228	disease	0.997
change	1.265	loss	0.997
cancer	1.306	infection	0.996
destruction	1.315	ton	0.995
disease	1.342	cancer	0.994

Cluster 38
size = 41.02574
 $\beta = 4.253067$

most likely verbs	
contain	0.081
produce	0.077
include	0.057
form	0.051
absorb	0.032

closest nouns		members	
salt	0.894	hormone	0.995
material	1.019	neutron	0.970
substance	1.047	radiation	0.960
dioxide	1.075	dioxide	0.942
chemical	1.139	light	0.939
oil	1.147	hair	0.925
enzyme	1.154	enzyme	0.893
molecule	1.168	ray	0.892
iron	1.179	moisture	0.878
protein	1.181	mineral	0.863

Cluster 39
size = 37.01715
 $\beta = 4.253067$

most likely verbs	
become	0.030
include	0.028
allow	0.026
enable	0.025
make	0.020

closest nouns		members	
people	1.242	slavery	0.992
animal	1.404	prey	0.991
person	1.413	insect	0.984
fish	1.458	user	0.982
modern	1.465	scientist	0.981
number	1.466	human	0.970
worker	1.468	student	0.958
plant	1.486	fish	0.945
student	1.511	host	0.944
artist	1.563	other	0.917

Cluster 40
size = 34.94247
 $\beta = 4.253067$

most likely verbs	
serve	0.045
use	0.034
become	0.033
include	0.019
produce	0.018

closest nouns		members	
dance	1.414	purpose	0.998
system	1.418	briefly	0.997
combination	1.421	victim	0.996
state	1.425	term	0.986
number	1.427	today	0.984
modern	1.461	sentence	0.973
variety	1.516	plot	0.927
tool	1.568	hand	0.927
structure	1.568	fire	0.865
medium	1.616	slave	0.701

Cluster 41
size = 31.9455
 $\beta = 4.253067$

most likely verbs	
take	0.251
assume	0.064
hold	0.036
win	0.023
become	0.022

closest nouns		members	
possession	0.936	throne	1.000
command	0.952	command	1.000
hold	0.956	place	0.999
form	0.983	office	0.999
control	1.033	responsibility	0.998
title	1.050	possession	0.998
advantage	1.063	advantage	0.997
shape	1.108	leadership	0.997
turn	1.140	step	0.996
step	1.163	root	0.995

Cluster 42
size = 36.45391
 $\beta = 4.253067$

most likely verbs	
provide	0.141
give	0.117
find	0.050
receive	0.045
seek	0.026

closest nouns		members	
impetus	0.838	impetus	0.998
explanation	0.948	rise	0.998
assistance	0.964	birth	0.994
protection	0.996	protection	0.992
insight	1.059	inspiration	0.990
representation	1.077	way	0.990
relief	1.167	access	0.987
evidence	1.199	expression	0.985
instruction	1.200	opportunity	0.981
access	1.200	employment	0.978

Cluster 43
size = 40.83409
 $\beta = 4.253067$

most likely verbs	
increase	0.063
improve	0.029
set	0.027
produce	0.025
reduce	0.024

closest nouns		members	
efficiency	1.339	foot	0.994
strength	1.348	efficiency	0.991
rate	1.351	scope	0.971
yield	1.382	supply	0.954
output	1.385	tone	0.943
content	1.423	potential	0.936
number	1.438	standard	0.928
voltage	1.454	voltage	0.927
ratio	1.458	quality	0.908
concentration	1.474	capacity	0.902

Cluster 44
size = 31.24072
 $\beta = 4.253067$

most likely verbs	
change	0.042
mark	0.037
place	0.034
determine	0.022
bring	0.022

closest nouns		members	
value	1.455	emphasis	0.999
direction	1.530	restriction	0.994
character	1.550	end	0.993
location	1.555	burden	0.991
color	1.579	direction	0.989
stress	1.583	stress	0.986
structure	1.589	beginning	0.985
number	1.682	limitation	0.974
concentration	1.735	trend	0.970
burden	1.741	side	0.963

Cluster 45
size = 42.20947
 $\beta = 4.253067$

most likely verbs	
show	0.070
reflect	0.038
express	0.033
reveal	0.017
share	0.017

closest nouns		members	
desire	1.168	influence	0.996
adaptation	1.219	concern	0.992
tendency	1.266	equation	0.983
diversity	1.302	sign	0.982
difference	1.403	tendency	0.979
relationship	1.422	gift	0.978
gift	1.494	existence	0.978
concern	1.496	emotion	0.978
awareness	1.509	desire	0.975
characteristic	1.518	talent	0.953

Cluster 47
size = 27.44644
 $\beta = 4.253067$

most likely verbs	
lead	0.111
join	0.102
form	0.066
succeed	0.030
send	0.022

closest nouns		members	
team	1.069	brother	1.000
army	1.145	rebellion	1.000
coalition	1.188	revolt	0.999
party	1.191	expedition	0.999
band	1.214	uprising	0.999
navy	1.254	father	0.998
league	1.268	faculty	0.998
union	1.333	opposition	0.996
force	1.339	army	0.995
uprising	1.389	staff	0.994

Cluster 49
size = 30.64602
 $\beta = 4.253067$

most likely verbs	
write	0.285
include	0.078
publish	0.072
produce	0.028
compose	0.023

closest nouns		members	
essay	0.574	biography	1.000
treatise	0.598	treatise	1.000
novel	0.672	novel	0.999
poem	0.677	journal	0.999
book	0.719	autobiography	0.999
biography	0.755	story	0.998
comedy	0.759	poetry	0.997
fiction	0.786	article	0.996
autobiography	0.821	verse	0.996
poetry	0.832	fiction	0.994

Cluster 46
size = 35.98153
 $\beta = 4.253067$

most likely verbs	
study	0.093
develop	0.076
teach	0.033
use	0.024
practice	0.023

closest nouns		members	
philosophy	1.190	medicine	1.000
science	1.197	theology	0.999
mathematic	1.205	philosophy	0.997
method	1.240	mathematic	0.996
architecture	1.247	architecture	0.996
theology	1.288	agriculture	0.996
concept	1.336	law	0.991
doctrine	1.337	piano	0.988
theory	1.396	theory	0.982
idea	1.406	concept	0.979

Cluster 48
size = 39.12344
 $\beta = 4.253067$

most likely verbs	
build	0.084
establish	0.078
found	0.046
become	0.042
include	0.038

closest nouns		members	
school	1.019	fort	0.996
community	1.057	mission	0.996
university	1.133	dynasty	0.994
house	1.172	nest	0.994
church	1.188	settlement	0.990
colony	1.202	college	0.987
temple	1.215	school	0.985
factory	1.260	regime	0.984
network	1.260	university	0.982
institution	1.294	house	0.981

Cluster 50
size = 38.71117
 $\beta = 4.253067$

most likely verbs	
include	0.117
contain	0.047
produce	0.044
make	0.028
use	0.028

closest nouns		members	
number	1.113	dialect	1.000
chemical	1.142	language	0.949
work	1.236	museum	0.900
modern	1.251	note	0.898
painting	1.262	collection	0.881
material	1.277	ballet	0.836
series	1.283	statue	0.818
piece	1.286	letter	0.737
paper	1.305	painting	0.721
variety	1.306	processing	0.719

Cluster 51
size = 30.14763
 $\beta = 4.253067$

most likely verbs	
make	0.067
attract	0.058
sign	0.028
involve	0.024
encourage	0.021

closest nouns		members	
conversion	1.407	treaty	0.999
improvement	1.426	agreement	0.999
addition	1.448	tourist	0.997
use	1.449	visitor	0.996
migration	1.450	contract	0.977
investment	1.452	female	0.970
removal	1.483	return	0.965
recovery	1.486	peace	0.961
establishment	1.496	revival	0.960
development	1.558	marriage	0.934

Cluster 53
size = 30.76758
 $\beta = 4.253067$

most likely verbs	
win	0.175
receive	0.152
play	0.035
obtain	0.024
earn	0.020

closest nouns		members	
award	0.836	game	0.999
nomination	0.864	vote	0.998
acclaim	0.972	role	0.996
approval	1.034	battle	0.996
prize	1.045	commission	0.995
medal	1.111	award	0.994
charter	1.179	allegiance	0.993
vote	1.326	doctorate	0.991
recognition	1.330	patent	0.989
doctorate	1.346	charter	0.988

Cluster 52
size = 33.81852
 $\beta = 4.253067$

most likely verbs	
make	0.395
earn	0.015
leave	0.014
involve	0.010
include	0.010

closest nouns		members	
voyage	0.616	debut	0.999
trip	0.697	contribution	0.999
contribution	0.770	trip	0.995
debut	0.776	fortune	0.995
loan	0.887	voyage	0.993
progress	0.927	loan	0.992
tour	0.974	tour	0.990
observation	1.000	living	0.988
attempt	1.007	attempt	0.988
concession	1.025	decision	0.986

Cluster 54
size = 32.29652
 $\beta = 4.253067$

most likely verbs	
achieve	0.103
gain	0.101
win	0.061
maintain	0.036
enjoy	0.031

closest nouns		members	
renown	0.813	prominence	1.000
popularity	0.823	fame	0.999
fame	0.915	popularity	0.999
prominence	0.931	success	0.998
recognition	1.087	independence	0.998
autonomy	1.148	renown	0.998
acceptance	1.159	reputation	0.994
status	1.193	acceptance	0.991
independence	1.193	goal	0.991
confidence	1.317	momentum	0.988

B Clusters from 1988 AP Newswire

The listing below shows in more detail the 25 leaf clusters for Figure 3.

Cluster 17 size = 53.08012 $\beta = 3.870965$	<table border="1"> <thead> <tr><th colspan="2">most likely verbs</th></tr> </thead> <tbody> <tr><td>make</td><td>0.021</td></tr> <tr><td>continue</td><td>0.018</td></tr> <tr><td>end</td><td>0.017</td></tr> <tr><td>follow</td><td>0.014</td></tr> <tr><td>stage</td><td>0.013</td></tr> </tbody> </table>	most likely verbs		make	0.021	continue	0.018	end	0.017	follow	0.014	stage	0.013	Cluster 20 size = 50.75491 $\beta = 3.870965$	<table border="1"> <thead> <tr><th colspan="2">most likely verbs</th></tr> </thead> <tbody> <tr><td>use</td><td>0.028</td></tr> <tr><td>carry</td><td>0.022</td></tr> <tr><td>make</td><td>0.019</td></tr> <tr><td>sell</td><td>0.016</td></tr> <tr><td>burn</td><td>0.016</td></tr> </tbody> </table>	most likely verbs		use	0.028	carry	0.022	make	0.019	sell	0.016	burn	0.016																																																																
most likely verbs																																																																																											
make	0.021																																																																																										
continue	0.018																																																																																										
end	0.017																																																																																										
follow	0.014																																																																																										
stage	0.013																																																																																										
most likely verbs																																																																																											
use	0.028																																																																																										
carry	0.022																																																																																										
make	0.019																																																																																										
sell	0.016																																																																																										
burn	0.016																																																																																										
<table border="1"> <thead> <tr><th colspan="2">closest nouns</th></tr> </thead> <tbody> <tr><td>protest</td><td>1.157</td></tr> <tr><td>effort</td><td>1.239</td></tr> <tr><td>strike</td><td>1.245</td></tr> <tr><td>use</td><td>1.264</td></tr> <tr><td>sale</td><td>1.272</td></tr> <tr><td>violence</td><td>1.284</td></tr> <tr><td>attack</td><td>1.334</td></tr> <tr><td>fight</td><td>1.341</td></tr> <tr><td>demonstration</td><td>1.348</td></tr> <tr><td>campaign</td><td>1.351</td></tr> </tbody> </table>	closest nouns		protest	1.157	effort	1.239	strike	1.245	use	1.264	sale	1.272	violence	1.284	attack	1.334	fight	1.341	demonstration	1.348	campaign	1.351	<table border="1"> <thead> <tr><th colspan="2">members</th></tr> </thead> <tbody> <tr><td>uprising</td><td>0.991</td></tr> <tr><td>terrorism</td><td>0.990</td></tr> <tr><td>riot</td><td>0.985</td></tr> <tr><td>blaze</td><td>0.982</td></tr> <tr><td>protest</td><td>0.957</td></tr> <tr><td>war</td><td>0.948</td></tr> <tr><td>violence</td><td>0.945</td></tr> <tr><td>unrest</td><td>0.942</td></tr> <tr><td>coup</td><td>0.933</td></tr> <tr><td>struggle</td><td>0.928</td></tr> </tbody> </table>	members		uprising	0.991	terrorism	0.990	riot	0.985	blaze	0.982	protest	0.957	war	0.948	violence	0.945	unrest	0.942	coup	0.933	struggle	0.928	<table border="1"> <thead> <tr><th colspan="2">closest nouns</th></tr> </thead> <tbody> <tr><td>weapon</td><td>1.194</td></tr> <tr><td>equipment</td><td>1.371</td></tr> <tr><td>material</td><td>1.386</td></tr> <tr><td>product</td><td>1.421</td></tr> <tr><td>food</td><td>1.439</td></tr> <tr><td>arm</td><td>1.475</td></tr> <tr><td>gun</td><td>1.479</td></tr> <tr><td>drug</td><td>1.495</td></tr> <tr><td>ton</td><td>1.499</td></tr> <tr><td>item</td><td>1.514</td></tr> </tbody> </table>	closest nouns		weapon	1.194	equipment	1.371	material	1.386	product	1.421	food	1.439	arm	1.475	gun	1.479	drug	1.495	ton	1.499	item	1.514	<table border="1"> <thead> <tr><th colspan="2">members</th></tr> </thead> <tbody> <tr><td>banner</td><td>0.987</td></tr> <tr><td>tire</td><td>0.977</td></tr> <tr><td>flag</td><td>0.967</td></tr> <tr><td>telephone</td><td>0.965</td></tr> <tr><td>phone</td><td>0.963</td></tr> <tr><td>machine</td><td>0.962</td></tr> <tr><td>oil</td><td>0.961</td></tr> <tr><td>hole</td><td>0.960</td></tr> <tr><td>satellite</td><td>0.935</td></tr> <tr><td>finger</td><td>0.891</td></tr> </tbody> </table>	members		banner	0.987	tire	0.977	flag	0.967	telephone	0.965	phone	0.963	machine	0.962	oil	0.961	hole	0.960	satellite	0.935	finger	0.891
closest nouns																																																																																											
protest	1.157																																																																																										
effort	1.239																																																																																										
strike	1.245																																																																																										
use	1.264																																																																																										
sale	1.272																																																																																										
violence	1.284																																																																																										
attack	1.334																																																																																										
fight	1.341																																																																																										
demonstration	1.348																																																																																										
campaign	1.351																																																																																										
members																																																																																											
uprising	0.991																																																																																										
terrorism	0.990																																																																																										
riot	0.985																																																																																										
blaze	0.982																																																																																										
protest	0.957																																																																																										
war	0.948																																																																																										
violence	0.945																																																																																										
unrest	0.942																																																																																										
coup	0.933																																																																																										
struggle	0.928																																																																																										
closest nouns																																																																																											
weapon	1.194																																																																																										
equipment	1.371																																																																																										
material	1.386																																																																																										
product	1.421																																																																																										
food	1.439																																																																																										
arm	1.475																																																																																										
gun	1.479																																																																																										
drug	1.495																																																																																										
ton	1.499																																																																																										
item	1.514																																																																																										
members																																																																																											
banner	0.987																																																																																										
tire	0.977																																																																																										
flag	0.967																																																																																										
telephone	0.965																																																																																										
phone	0.963																																																																																										
machine	0.962																																																																																										
oil	0.961																																																																																										
hole	0.960																																																																																										
satellite	0.935																																																																																										
finger	0.891																																																																																										
Cluster 24 size = 57.6904 $\beta = 3.870965$	<table border="1"> <thead> <tr><th colspan="2">most likely verbs</th></tr> </thead> <tbody> <tr><td>give</td><td>0.069</td></tr> <tr><td>receive</td><td>0.035</td></tr> <tr><td>get</td><td>0.034</td></tr> <tr><td>seek</td><td>0.026</td></tr> <tr><td>provide</td><td>0.025</td></tr> </tbody> </table>	most likely verbs		give	0.069	receive	0.035	get	0.034	seek	0.026	provide	0.025	Cluster 27 size = 33.35255 $\beta = 3.88911$	<table border="1"> <thead> <tr><th colspan="2">most likely verbs</th></tr> </thead> <tbody> <tr><td>make</td><td>0.287</td></tr> <tr><td>force</td><td>0.023</td></tr> <tr><td>announce</td><td>0.017</td></tr> <tr><td>include</td><td>0.017</td></tr> <tr><td>follow</td><td>0.015</td></tr> </tbody> </table>	most likely verbs		make	0.287	force	0.023	announce	0.017	include	0.017	follow	0.015																																																																
most likely verbs																																																																																											
give	0.069																																																																																										
receive	0.035																																																																																										
get	0.034																																																																																										
seek	0.026																																																																																										
provide	0.025																																																																																										
most likely verbs																																																																																											
make	0.287																																																																																										
force	0.023																																																																																										
announce	0.017																																																																																										
include	0.017																																																																																										
follow	0.015																																																																																										
<table border="1"> <thead> <tr><th colspan="2">closest nouns</th></tr> </thead> <tbody> <tr><td>information</td><td>1.193</td></tr> <tr><td>assurance</td><td>1.274</td></tr> <tr><td>aid</td><td>1.321</td></tr> <tr><td>notice</td><td>1.341</td></tr> <tr><td>answer</td><td>1.345</td></tr> <tr><td>report</td><td>1.359</td></tr> <tr><td>coverage</td><td>1.381</td></tr> <tr><td>copy</td><td>1.398</td></tr> <tr><td>treatment</td><td>1.406</td></tr> <tr><td>protection</td><td>1.430</td></tr> </tbody> </table>	closest nouns		information	1.193	assurance	1.274	aid	1.321	notice	1.341	answer	1.345	report	1.359	coverage	1.381	copy	1.398	treatment	1.406	protection	1.430	<table border="1"> <thead> <tr><th colspan="2">members</th></tr> </thead> <tbody> <tr><td>warrant</td><td>0.978</td></tr> <tr><td>injunction</td><td>0.939</td></tr> <tr><td>notice</td><td>0.927</td></tr> <tr><td>sentence</td><td>0.894</td></tr> <tr><td>warning</td><td>0.890</td></tr> <tr><td>letter</td><td>0.875</td></tr> <tr><td>purpose</td><td>0.875</td></tr> <tr><td>permit</td><td>0.870</td></tr> <tr><td>signal</td><td>0.857</td></tr> <tr><td>visa</td><td>0.857</td></tr> </tbody> </table>	members		warrant	0.978	injunction	0.939	notice	0.927	sentence	0.894	warning	0.890	letter	0.875	purpose	0.875	permit	0.870	signal	0.857	visa	0.857	<table border="1"> <thead> <tr><th colspan="2">closest nouns</th></tr> </thead> <tbody> <tr><td>recommendation</td><td>0.972</td></tr> <tr><td>decision</td><td>0.973</td></tr> <tr><td>contribution</td><td>1.036</td></tr> <tr><td>announcement</td><td>1.049</td></tr> <tr><td>concession</td><td>1.058</td></tr> <tr><td>choice</td><td>1.118</td></tr> <tr><td>mention</td><td>1.134</td></tr> <tr><td>change</td><td>1.145</td></tr> <tr><td>progress</td><td>1.147</td></tr> <tr><td>appearance</td><td>1.165</td></tr> </tbody> </table>	closest nouns		recommendation	0.972	decision	0.973	contribution	1.036	announcement	1.049	concession	1.058	choice	1.118	mention	1.134	change	1.145	progress	1.147	appearance	1.165	<table border="1"> <thead> <tr><th colspan="2">members</th></tr> </thead> <tbody> <tr><td>debut</td><td>0.995</td></tr> <tr><td>mistake</td><td>0.993</td></tr> <tr><td>mention</td><td>0.990</td></tr> <tr><td>landing</td><td>0.974</td></tr> <tr><td>reference</td><td>0.966</td></tr> <tr><td>comment</td><td>0.958</td></tr> <tr><td>plea</td><td>0.948</td></tr> <tr><td>evacuation</td><td>0.942</td></tr> <tr><td>announcement</td><td>0.933</td></tr> <tr><td>stop</td><td>0.927</td></tr> </tbody> </table>	members		debut	0.995	mistake	0.993	mention	0.990	landing	0.974	reference	0.966	comment	0.958	plea	0.948	evacuation	0.942	announcement	0.933	stop	0.927
closest nouns																																																																																											
information	1.193																																																																																										
assurance	1.274																																																																																										
aid	1.321																																																																																										
notice	1.341																																																																																										
answer	1.345																																																																																										
report	1.359																																																																																										
coverage	1.381																																																																																										
copy	1.398																																																																																										
treatment	1.406																																																																																										
protection	1.430																																																																																										
members																																																																																											
warrant	0.978																																																																																										
injunction	0.939																																																																																										
notice	0.927																																																																																										
sentence	0.894																																																																																										
warning	0.890																																																																																										
letter	0.875																																																																																										
purpose	0.875																																																																																										
permit	0.870																																																																																										
signal	0.857																																																																																										
visa	0.857																																																																																										
closest nouns																																																																																											
recommendation	0.972																																																																																										
decision	0.973																																																																																										
contribution	1.036																																																																																										
announcement	1.049																																																																																										
concession	1.058																																																																																										
choice	1.118																																																																																										
mention	1.134																																																																																										
change	1.145																																																																																										
progress	1.147																																																																																										
appearance	1.165																																																																																										
members																																																																																											
debut	0.995																																																																																										
mistake	0.993																																																																																										
mention	0.990																																																																																										
landing	0.974																																																																																										
reference	0.966																																																																																										
comment	0.958																																																																																										
plea	0.948																																																																																										
evacuation	0.942																																																																																										
announcement	0.933																																																																																										
stop	0.927																																																																																										

Cluster 28
size = 35.25271
 $\beta = 3.88911$

most likely verbs	
file	0.053
make	0.053
sign	0.037
reject	0.034
reach	0.027

closest nouns		members	
proposal	0.964	fun	0.999
agreement	1.171	suit	0.995
appeal	1.198	motion	0.992
pact	1.214	petition	0.989
request	1.248	lawsuit	0.989
application	1.248	verdict	0.944
measure	1.264	accord	0.939
recommendation	1.288	treaty	0.938
claim	1.297	conclusion	0.932
declaration	1.363	complaint	0.928

Cluster 30
size = 39.467
 $\beta = 3.88911$

most likely verbs	
play	0.029
make	0.026
take	0.019
include	0.015
get	0.015

closest nouns		members	
year	1.170	pool	0.999
state	1.298	song	0.967
thing	1.331	role	0.957
program	1.339	sound	0.851
today	1.348	music	0.846
week	1.366	game	0.841
child	1.378	hit	0.814
music	1.388	politic	0.772
woman	1.427	character	0.738
company	1.460	conversation	0.724

Cluster 32
size = 26.85736
 $\beta = 3.88911$

most likely verbs	
fire	0.052
carry	0.039
break	0.031
lose	0.021
use	0.021

closest nouns		members	
gun	1.063	finger	1.000
weapon	1.337	banner	1.000
pistol	1.446	window	0.999
missile	1.484	glass	0.997
rifle	1.500	bullet	0.996
arm	1.555	flag	0.995
glass	1.613	rocket	0.994
rocket	1.735	pistol	0.989
flag	1.778	rifle	0.978
ton	1.844	satellite	0.974

Cluster 29
size = 34.94556
 $\beta = 3.88911$

most likely verbs	
hold	0.037
spend	0.033
tell	0.032
attend	0.031
begin	0.030

closest nouns		members	
week	0.632	slogan	1.000
year	0.670	convention	0.988
month	0.734	ceremony	0.979
day	0.795	dec.	0.969
today	0.885	feb.	0.960
meeting	1.094	night	0.957
weekend	1.163	contest	0.953
summer	1.194	oct.	0.953
hour	1.202	session	0.939
oct.	1.212	weekend	0.936

Cluster 31
size = 16.98393
 $\beta = 3.88911$

most likely verbs	
throw	0.085
hurl	0.072
run	0.067
use	0.039
smoke	0.030

closest nouns		members	
grenade	1.219	firebomb	1.000
bomb	1.262	stone	1.000
explosive	1.357	rock	1.000
rock	1.370	grenade	1.000
device	1.486	mate	0.999
stone	1.501	bomb	0.998
firebomb	1.518	egg	0.997
cocaine	1.525	marijuana	0.997
marijuana	1.678	gas	0.996
ring	1.800	cigarette	0.984

Cluster 33
size = 52.47398
 $\beta = 3.88911$

most likely verbs	
tell	0.109
become	0.038
include	0.024
ask	0.022
elect	0.016

closest nouns		members	
member	0.831	editor	0.998
president	0.874	analyst	0.995
official	0.909	director	0.992
leader	0.992	chairman	0.979
minister	1.042	reporter	0.968
attorney	1.046	mayor	0.952
judge	1.110	juror	0.946
senator	1.126	jury	0.935
group	1.129	minister	0.927
student	1.132	investigator	0.924

Cluster 34
size = 52.06246
 $\beta = 3.88911$

most likely verbs	
kill	0.037
tell	0.031
include	0.026
help	0.017
allow	0.015

closest nouns		members	
people	0.610	murderer	0.999
man	0.668	policeman	0.954
woman	0.711	civilian	0.945
member	0.766	protester	0.911
student	0.805	demonstrator	0.894
worker	0.942	dealer	0.879
child	0.972	voter	0.876
person	1.034	guerrilla	0.854
soldier	1.062	other	0.852
employee	1.112	youth	0.849

Cluster 36
size = 32.62717
 $\beta = 3.88911$

most likely verbs	
rise	0.029
sell	0.023
fall	0.020
buy	0.020
make	0.017

closest nouns		members	
amount	1.072	cent	1.000
number	1.127	yen	0.997
share	1.220	point	0.979
percent	1.235	percent	0.977
price	1.261	stock	0.970
level	1.269	basis	0.946
dollar	1.325	asset	0.820
money	1.346	currency	0.816
rate	1.374	stake	0.792
total	1.375	average	0.785

Cluster 38
size = 31.8062
 $\beta = 3.88911$

most likely verbs	
take	0.210
hold	0.027
make	0.022
get	0.021
give	0.019

closest nouns		members	
part	1.000	advantage	0.995
place	1.036	place	0.992
advantage	1.109	step	0.972
step	1.133	look	0.968
look	1.147	stand	0.967
position	1.160	turn	0.964
action	1.164	responsibility	0.899
break	1.204	refuge	0.888
turn	1.287	effect	0.875
care	1.299	post	0.828

Cluster 35
size = 35.47928
 $\beta = 3.88911$

most likely verbs	
reduce	0.044
increase	0.042
pay	0.031
raise	0.022
cut	0.021

closest nouns		members	
number	0.938	output	0.970
production	1.061	expense	0.953
cost	1.091	debt	0.946
rate	1.098	spending	0.938
amount	1.100	deficit	0.903
sale	1.212	earning	0.897
spending	1.217	size	0.874
price	1.248	cost	0.860
wage	1.259	income	0.856
level	1.268	tax	0.851

Cluster 37
size = 34.37653
 $\beta = 3.88911$

most likely verbs	
win	0.090
seek	0.069
give	0.040
grant	0.036
get	0.036

closest nouns		members	
nomination	1.180	praise	0.998
approval	1.192	re-election	0.995
permission	1.214	asylum	0.988
recognition	1.224	prize	0.951
support	1.251	nomination	0.949
vote	1.298	injunction	0.942
protection	1.315	stay	0.935
seat	1.327	visa	0.933
access	1.332	presidency	0.917
award	1.354	independence	0.917

Cluster 39
size = 27.53791
 $\beta = 3.88911$

most likely verbs	
raise	0.045
express	0.034
cast	0.028
increase	0.025
face	0.024

closest nouns		members	
fear	1.224	doubt	0.999
concern	1.280	ballot	0.999
number	1.459	obstacle	0.994
threat	1.477	fear	0.987
opposition	1.506	danger	0.986
demand	1.528	question	0.984
risk	1.528	tension	0.976
danger	1.533	objection	0.956
challenge	1.588	expectation	0.939
hope	1.602	barrier	0.918

Cluster 40
size = 34.86093
 $\beta = 3.88911$

most likely verbs	
express	0.021
make	0.019
announce	0.017
give	0.015
show	0.014

closest nouns		members	
number	1.330	intention	0.971
increase	1.372	belief	0.965
view	1.421	veto	0.963
interest	1.516	theme	0.947
state	1.528	identity	0.886
support	1.534	importance	0.853
policy	1.561	confidence	0.839
change	1.583	desire	0.816
position	1.604	spirit	0.715
commitment	1.606	trend	0.653

Cluster 42
size = 38.80711
 $\beta = 3.88911$

most likely verbs	
leave	0.032
close	0.029
occupy	0.028
build	0.022
enter	0.018

closest nouns		members	
area	1.023	territory	1.000
city	1.081	street	0.998
building	1.130	border	0.992
house	1.190	gap	0.991
town	1.262	hall	0.989
plant	1.278	door	0.980
store	1.318	zone	0.970
facility	1.333	shop	0.970
hotel	1.372	bridge	0.969
park	1.404	land	0.967

Cluster 44
size = 39.30971
 $\beta = 3.88911$

most likely verbs	
scatter	0.021
make	0.019
take	0.018
get	0.018
change	0.016

closest nouns		members	
year	1.158	shower	0.999
state	1.177	snow	0.986
company	1.277	wind	0.921
program	1.300	rain	0.873
child	1.306	mind	0.749
today	1.310	weather	0.413
nation	1.341	path	0.304
number	1.368	air	0.289
city	1.374	way	0.228
family	1.409	generation	0.225

Cluster 41
size = 43.38313
 $\beta = 3.88911$

most likely verbs	
include	0.015
tell	0.014
leave	0.012
use	0.012
lead	0.011

closest nouns		members	
company	1.031	block	0.993
state	1.104	band	0.898
group	1.107	boat	0.861
city	1.121	force	0.838
family	1.239	troop	0.780
nation	1.241	vessel	0.746
student	1.252	coalition	0.735
member	1.258	carrier	0.659
child	1.260	mile	0.551
man	1.286	ship	0.535

Cluster 43
size = 34.39544
 $\beta = 3.88911$

most likely verbs	
improve	0.032
acquire	0.027
endanger	0.018
make	0.016
create	0.015

closest nouns		members	
nation	1.333	syndrome	0.998
system	1.345	species	0.994
company	1.411	relation	0.977
program	1.423	safety	0.937
security	1.429	memory	0.924
relationship	1.469	atmosphere	0.907
state	1.490	health	0.895
number	1.527	environment	0.891
health	1.535	quality	0.888
stability	1.550	tie	0.886

Cluster 45
size = 25.05437
 $\beta = 3.88911$

most likely verbs	
commit	0.056
deny	0.031
allege	0.025
cause	0.024
attempt	0.024

closest nouns		members	
violence	1.185	suicide	1.000
abuse	1.227	crime	0.999
problem	1.282	fraud	0.998
violation	1.338	dispute	0.989
conflict	1.452	murder	0.988
fraud	1.465	error	0.986
killing	1.520	terrorism	0.984
unrest	1.533	disease	0.970
crime	1.583	rumor	0.963
crisis	1.623	abuse	0.961

Cluster 46
size = 24.29099
 $\beta = 3.88911$

most likely verbs	
suffer	0.073
cause	0.069
report	0.027
follow	0.025
arm	0.020

closest nouns		members	
loss	1.040	wound	1.000
failure	1.137	blow	1.000
drop	1.203	injury	0.999
accident	1.245	defeat	0.998
decline	1.328	casualty	0.997
crash	1.392	damage	0.992
injury	1.401	robbery	0.974
death	1.408	death	0.966
shortage	1.446	pain	0.960
attack	1.447	loss	0.944

Cluster 48
size = 31.86115
 $\beta = 3.88911$

most likely verbs	
impose	0.040
set	0.037
meet	0.033
violate	0.024
include	0.014

closest nouns		members	
rule	1.017	curfew	1.000
regulation	1.133	sanction	0.989
policy	1.185	obligation	0.983
ban	1.203	deadline	0.979
requirement	1.281	restriction	0.974
law	1.287	ban	0.965
standard	1.337	standard	0.959
restriction	1.362	requirement	0.956
measure	1.385	law	0.944
quota	1.399	quota	0.931

Cluster 47
size = 42.94112
 $\beta = 3.88911$

most likely verbs	
make	0.027
take	0.016
include	0.014
give	0.013
get	0.012

closest nouns		members	
program	1.019	budget	0.793
year	1.108	strategy	0.352
state	1.196	schedule	0.289
week	1.285	path	0.284
company	1.301	course	0.249
today	1.304	procedure	0.239
number	1.327	program	0.238
series	1.329	table	0.227
sale	1.338	process	0.224
change	1.345	venture	0.218