# A Methodical Approach to Word Class Formation Using Automatic Evaluation

**John Hughes and Eric Atwell**

Centre for Computer Analysis of Language and Speech
School of Computer Studies
Leeds University
Leeds LS2 9JT, UK

e-mail: john@scs.leeds.ac.uk
eric@scs.leeds.ac.uk
phone: +44 (0)532 335430
fax: +44 (0)532 335468

## Abstract

Automatic inference of a classification of words has been carried out by several researchers recently. Although they use a variety of methods they all exploit the statistical redundancy inherent in the structure of language to differentiate words; the assumption being that words of similar rôles occur in measurably similar contexts.

This paper describes a general method by which clustering schemes can be qualitatively compared. This allows a systematic approach to finding the best word class formation scheme to be adopted. The process by which words are automatically grouped into classes involves a number of decision points. These include: the contextual pattern in the language being measured; the metric by which words are compared according to the pattern; and the mechanism by which items judged to be similar are merged. Alternatives are presented for each of these factors. The experiments rated each combination so that the most successful approach can be found. Previously, researchers relied on a *looks-good-to-me* method of self evaluation to judge the quality of their derived word classifications. This paper directly compares some of their adopted approaches with alternative clustering schemes not previously attempted. This allows us to formally demonstrate when our approach to clustering is more successful. The evaluation method is also shown to be a valuable aid to highlighting approaches that are inefficient.

Amongst the patterns investigated were the morphological context supplied by the previous words. Bigram counts of the collocation of the words to be clustered with the last three letters of the word immediately before were found to be a remarkably good differentiation criteria. The evaluation method demonstrated that the context of the last three letters (which on average contain a lot of morphological information in English) is even better that the context supplied by using the whole of the previous word in collocation counts. Results such as this should prove useful to handwriting recognition research. The authors believe this method provides a sensible first step for handwriting recognition researchers who wish to use statistical models of language to aid the disambiguation process; proposed contextual models can be evaluated relative to previously investigated models to indicate the likely success rate of employing them. This allows a proposed poor disambiguation method to be ruled out early on and thus is a valuable aid to saving valuable time and resources.

We end by considering some further applications of automatic word class formation techniques. Although our experiments are exclusively with English corpus text, the general clustering and word-classifying algorithms should be applicable to text in other languages. This is likely to be particularly useful in development of linguistic engineering technologies for emerging nations and their new mother tongues, which have little or no computational linguistics resources or computational linguistics to "hand-craft" them.

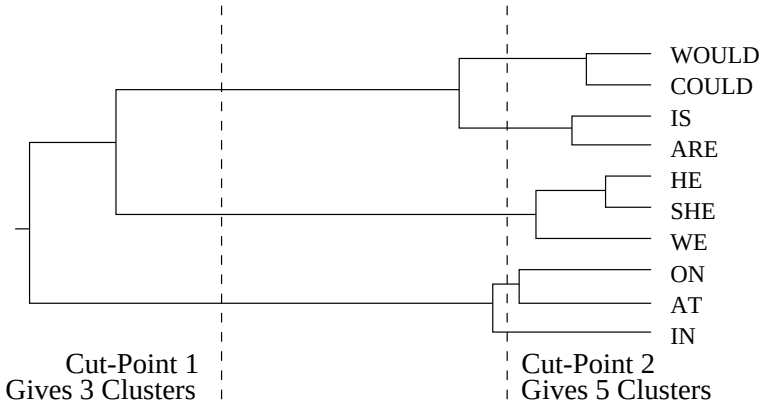
# 1 Introduction

Hierarchical clustering is a way to produce a taxonomic classification of items such that, for a given cut-off point, the cut-off groups contain homogenous objects whilst the groups are as heterogeneous amongst themselves as possible. The items must have initially been compared with each other in such a way as there is a standard measure of similarity between each pair. The process begins by finding the closest two items and replacing them by a measurement which represents the *union* of the two in some meaningful way. Then, the second closest pair of items are searched for. This second

group may consist of the first group merged with another item or it may consist of two new items. The items are collapsed in this way, iteratively, until all items become merged into the same group. As two items (or groups of items) merge a record is kept of their similarity and a dendrogram forms. The method is described in further detail in [Hughes 94], [Hughes and Atwell 93, 94].

There are several patterns in language that can provide a measure of similarity for words. $N$-gram counts of words have traditionally been the most commonly used measure but others such as the positional distribution of words might supply a useful context also. An example dendrogram is displayed in figure 1, below.

Figure 1: An Example Dendogram Showing Cut-Points




The cut-point lines are added for explanation (see section 2.2) but the remainder of the diagram is of the form automatically generated by the clustering program.

The choice of algorithm to calculate the distance of the two newly clustered items to the other items as well as the distance metric to initially compare vectors can have a profound influence on the clustering. Each combination of metric and clustering method was tried in the experiments to see which derived the strongest syntactic classification of words in comparison to an intuitive linguistic classification (by the evaluation mechanisms described below). Three metrics were considered here: Manhattan, Euclidean and Spearman Rank Correlation Coefficient. The latter follows the modified definition given by [Finch 93] so that our results can be directly compared. Likewise, the choice of clustering method can greatly alter the resultant dendrogram after clustering; eight methods were included in the experiments described here. [Zupan 82] gives iterative formulae for seven of the methods. The other clustering method uses the geometric centre of gravity to calculate the dissimilarity between the most recently joined pair and all other items.

# 2 Automatic Evaluation

The last essential part of the automatic word classification process is some means of rating the quality of the alternative clustering schema for their accuracy. Other word classification projects have failed to include this vital procedure.

## 2.1 *Looks Good to Me*

Evaluating a clustering is typically done by the programmer using a *looks good to me* approach. To an extent the programmer can feel how good one clustering is over an-

other because he/she has an intrinsic understanding of the processes that produced it. However, the programmer also has a vested interest in making his/her program look good. A more worthy evaluation can be done by an "independent" expert - in this case a linguist. It is rare to find one that has no bias in some way but the linguist's judgement based on experience must rate his/her appraisal above that of the programmer who has a vested interest to be seen to have done good work.

These evaluations are all done with some preconceived intuitive classification in mind. The actual question of what makes a good classification is not a simple one to answer. There are many alternatives and deciding which is superior comes down to personal judgement. Two rival clusterings may produce one winner when judged by one expert linguist but the other according to a different linguist's intuition. The linguist's intuition does not involve quantitative, measurable criteria, only qualitative overall impressions.

The *looks good to me* approach may be fine if the aim is to merely demonstrate that patterns in text can classify words. This in itself is a laudable aim but if the best possible classification is desired then some way of comparing clustering schemes is needed.

## 2.2 The LOB Benchmark Clustering

If it is accepted that a classification should conform closely to a syntactic intuitive one then there is a way it can be evaluated automatically thus resolving the problems of subjectivity amongst programmers and expert linguists. A *benchmark* classification can be derived which requires no input from the programmer nor a linguist but can be created empirically using a tagged corpus. A benchmark was derived from the tagged LOB corpus using a reduced tag-set [Hughes 94]. The novelty of the technique is that it yields a quantitative comparison against an existing corpus-based benchmark. In principle the algorithm could equally be applied using another tagged corpus as a base.

The evaluation tool works by cutting the dendrogram at a certain point to produce a number of clusters. The members of the clusters can then be examined to see how they are tagged in the reduced LOB tag-set. A score can be calculated by classifying each group as the most common type amongst its members and counting up how many members conform to this type.

The cut-point chosen will have a bearing on this process. The dendrogram can be cut at any point to produce any number of clusters. If the dendrogram is cut at the root there will be only one cluster containing all the items. If the dendrogram is cut at the leaves there are as many clusters as there are objects to begin with. Figure 1 shows the dendrogram being cut at two points to produce 3 and 5 clusters. The first cut divides the clustering into three groups that match intuitive expectations.

The benchmark consists of 19 'reduced' tags such as *noun*, *past tense verb* and *cardinal number*. An ideal clustering would match the benchmark ideally and would thus have 19 clusters. The dendrogram is cut at the point that produces 25 clusters which is very close to the ideal of 19 but still allows a little leeway. Deciding where to cut the dendrogram is obviously fairly ad hoc and other researchers in this area have skirted the issue and arbitrarily chosen a cut point that produces a relatively large number of groups which are likely to be homogenous because they do not have many members. However, some of the experiments avoid the cut-point issue altogether by cutting the dendrogram at many points throughout its width. Two rival clustering schemes can then be contrasted by plotting graphs of the evaluations throughout the range of cut-points

(see figure 2).

### 2.3 Automatically Evaluating Any Given Clustering

An alternative evaluation scheme does not use the benchmark but instead looks at the tagged LOB corpus to find how every word in the clustering is tagged. The rules follow from the benchmark used in the LOB experiments. Each word is compared with the LOB corpus to examine how it is tagged most often. The scoring regime follows that for the benchmark clustering.

The evaluator written for the 2000 word experiment also evaluates each group. An example of one of the least consistent groups (on the whole most groups are much more consistent than this as will be shown in the examples given later) looks like this:

```
13) NOUN    85.3261%
.HALF     *CHILD    *FLAME     .SET      *FIGURE     POSTING    RESUME    DIE
 ROUND    *CAT      *DREAM     *SIGN     *ANSWER    *COMMENT   *DRINK    *SLEEP
*CHIP     *DOG      *REQUEST   *WASTE    .OFFER     *REPLY     *SWING    .LIE
*BOY      *KID      *SURPRISE   e-mail   .GAIN      *DEAL      *DRESS    .FALL
*DOCTOR   *STEP     *BRAND      email    *PURCHASE  *CONTACT   *DANCE    *VOTE
*BABY     .DAMN      MIX       *MAIL     *POST      *TOUCH     *TRADE    *WORK
```

If a word was tagged most frequently in LOB the same way as the tag assigned to its cluster (such as the majority of words in the example) then it was marked with a "*". If, instead, the second, third or fourth most common tag for the word in LOB matched its cluster's assigned tag (such as the words *half*, *damn*, *set*, *offer gain*, *lie* or *fall* in the example) then that word is marked with a ".". Words that do not match up (such as the words *round mix* or *die* in the example) aren't annotated at all. The words that are not present in LOB (*e-mail* and *email*) are printed in lower case whilst the recognised words are converted to upper case. The unknown words aren't included in any of the evaluation counts. A score out of 100 is calculated for each cluster using the same scoring methods for calculating an overall score. The example group was declared a **NOUN** group by the evaluator with approximately 85% accuracy.

## 3 Results

This section briefly records some of the results of various clustering schemes applied to some of the patterns in English language. The first set of experiments were carried out on a sample set of the 200 most frequent words in the LOB corpus as they appear in the untagged LOB corpus. The evaluation tool demonstrates which clustering scheme produces groupings most in line with intuitive expectations and this scheme is used in experiments to cluster much larger groups of words.

### 3.1 Finding the Best Clustering Method

Table 1, below, contrasts the results for three distributional patterns formed by the position of a word in a sentence and two types of bigram counts. Normalized vectors were derived from statistics sampling the three patterns. Each combination of three metrics and eight clustering techniques were used to cluster the vectors (except for some of the third set of experiments - marked with a '—' - were results of certain

combinations had already proved themselves not worthy of further investigation). The resultant dendrograms were evaluated, for the cut-off point where there were 25 clusters, against the benchmark clustering. Each cell in the table, then, shows three figures: the first, on the left, for the combination of metric and clustering method ran on the statistics derived from sentence position distribution; the second, in the centre, from the distribution of immediate neighbour bigrams; the third, on the right, from the n–2, n–1, n+1, n+2 bigram distribution.

The evaluations reveal that the context implied by sentence position distribution provides a poor representation of the syntactic rôle of the 200 words. The highest scoring combination consisting of the Euclidean metric and Ward's clustering method was only judged to be about 45% correct. The second set of experiments, on bigram counts of the 200 most frequent words appearing immediately before of after a target set of the most frequent 101 lexical items, scored a great deal better than for the sentence position distribution. The highest scoring combination, Manhattan metric and Ward's clustering method scored 76%. The poor relative performance of sentence position distribution as a context measure meant it was not investigated further. However, there was clearly scope to investigate bigrams further. A third set of experiments, this time on just the best performing clustering schemes from the earlier experiments were carried out for bigrams covering the closest two neighbours on either side. These results are detailed on the right of the cells in Table 1.

**Table 1: Evaluations for the Following Distributions:**
**(left) the position of a word in a sentence**
**(centre) bigrams for positions n–1, n+1**
**(right) bigrams for positions n–2, n–1, n+1, n+2**

| Metric | Single Linkage | Complete Linkage | Group Average | Weighted Grp. Ave. |
|---|---|---|---|---|
| Manhattan | 25 38 — | 42 69 75 | 38 72 70 | 40 73 74 |
| Euclidian | 29 31 — | 42 60 — | 37 46 — | 41 50 — |
| Spearman Rank | 23 29 — | 41 75 76 | 36 74 69 | 41 70 71 |

| Metric | Median | Centroid | Centre of Gravity | Ward's Method |
|---|---|---|---|---|
| Manhattan | 27 29 — | 23 26 — | 27 42 — | 43 76 79 |
| Euclidian | 28 31 — | 27 32 — | 37 45 — | 45 64 — |
| Spearman Rank | 27 26 — | 28 26 — | 32 67 — | 42 74 77 |

### 3.1.1 Evaluating the Context Supplied by Bigrams

Intuitively we would imagine that the context provided by the nearest words would be more valuable than from words further away. This can be verified by clustering using just the bigram counts for certain relative positions to the target words.

**Table 2: Results for Experiments on Each of the Six Bigram Positions**

| Bigram Position | n–3 | n–2 | n–1 | n+1 | n+2 | n+3 |
|---|---|---|---|---|---|---|
| Score | 55 | 60 | 69 | 61 | 53 | 50 |

The results listed in table 2 confirm these expectations. The immediate neighbours supply a better context than those further away and the best context of all is supplied by the $n$–1 bigrams.

### 3.1.2 Morphological Context

To investigate the power of morphological context a further experiment was initiated in which the context was calculated using just the previous words' last three letters. The hundred most frequent three letter word-endings were found. Then bigram counts were calculated of the number of times each of the two hundred words to be clustered appeared immediately after a word which ended with one of the items in the reference set. Using this as the sole context a clustering was made which was evaluated to around 73% compared with about 69% for the evaluation for the clustering using the whole of the previous word as the context. This implies that, in English, the rôle of a word can be predicted with a high degree of accuracy using the endings of words.

This result should be of particular interest to researchers using parts of words as context for such tasks as handwriting recognition. [Hanlon and Boyle 92], for example, are using the endings of words to suggest the likely syntactic rôle for handwritten words that have more than one possible derivation according to a handwriting recogniser. Thus the alternatives which would seem unlikely in the context of the word endings can be given a lower rank than the alternatives that fit more naturally with the syntactic context.

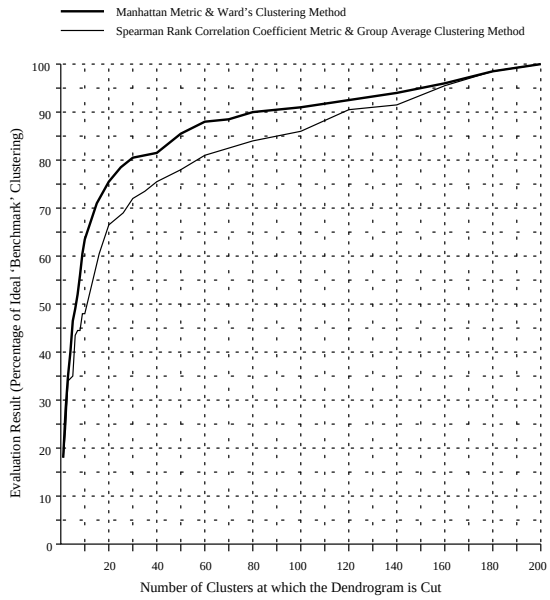



Figure 2: Evaluation Graphs for Two Alternative Clustering Schemes

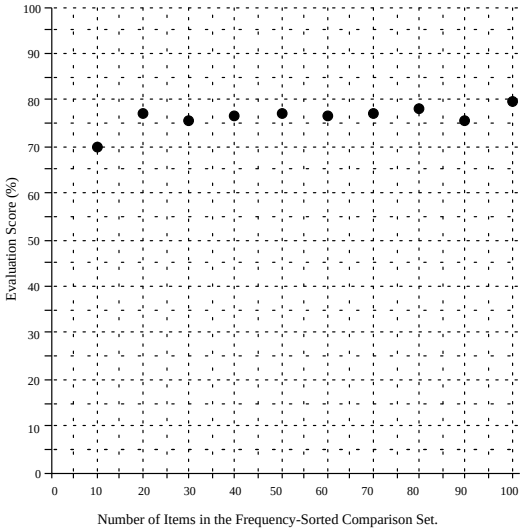



Figure 3: Evaluations for Comparison Sets of Varying Size

### 3.1.3 Varying the Cut-Points in the Dendrograms

One factor of the experimental procedure which may lead to false bias was the point at which the dendrogram was cut to form $n$ clusters. Any bias due to the high dendrogram cut-off point used in the evaluator can be side-stepped if a graph is plotted for evaluations over a range of values. Figure 2 compares the highest scoring combination from our experiments, Manhattan metric and Ward's clustering method, with the combination that Finch believed to work best in his experiments, the Spearman Rank Correlation Coefficient metric and the Group Average clustering method. Clearly the combination of Manhattan metric and Ward's clustering method consistently outperforms the rival

clustering scheme when evaluated for cut points throughout the dendrogram.

### 3.1.4 Varying the Size of the Comparison Set

To investigate the effect of the number of items in the comparison set ten experiments were carried out, each using ten more items than the previous one with the items being added in order of frequency. The results of these ten experiments are plotted in Figure 3. Just the ten most frequent lexical items lead to an evaluation of almost 70%. Adding in more and more items into the comparison set makes no significant difference to the quality of the clustering as measured by the evaluation tool. The reason the expressive power of the most frequent lexical items is so good is because they are mainly function words. [Powers 92] suggests that as these words are relatively unaffected by domain they act as markers for other words, hence indicating the categories of those words. In Schütze's experiments to cluster 5000 words he used the context of bigram counts in the positions $n$-2, $n$-1, $n$+1 and $n$+2 as they co-occurred with the same 5000 words [Schütze 93]. As the best contextual information seems to be provided by the function words - which make up the major part of the most frequent words in the corpus - it seems wasteful on resources to have such a large comparison set.

# 4 A Clustering of 2000 words

Now that the factors leading to a good clustering of words had been investigated we could select the best clustering scheme and use it to cluster a much larger set of words. The distributional context of $n-2$, $n-1$, $n+1$, $n+2$ bigram counts, the Manhattan metric and Ward's clustering algorithm were used to cluster 2000 words When scored according to the evaluator the results were demonstrably good. For corpora of size 16 million and 35 million words the evaluations are very similar. When the dendrograms are cut at the point where there are 25 clusters (a very tightly constrained set for 2000 words) both scores are in the region of 80%. This implies that the corpus of 16 million words (a *third* the size of Finch's corpus) is representative of the bigram distribution and there is little to gain from using larger corpora. The large-scale clustering was shown to not only group items of similar syntax but also to partially cluster items on their semantic or morphological similarity. When the dendrogram was cut to make 100 clusters the groups listed on the next page were amongst the cut-off clusters. The numbers in the list are labels to identify the location of the groups in the dendrogram.

20 Day Night Afternoon Morning Summer Weekend Century Season Month Week Year
22 Days Hours Minutes Weeks Months Years
28 Feet Hands Fingers Eyes Legs Clothes Hair Arms Teeth Mind Opinion Chest Mouth Ass Breath Tongue Foot Arm Shoulder Face Head Heart Memory Name Voice
29 Brother Sister Father Mother Daughter Son Mom Husband Wife
36 Australia Canada America Europe Cuba Lebanon California Boston Chicago Vietnam
75 Said Says Knows Feels Believes Thinks Assumed Believed Claimed Meant Stated Suggested Felt Knew Realized Figured Thought
79 Adding Allowing Causing Leaving Letting Bringing Giving Putting Sending Finding Keeping Having Buying Making Taking Using
82 David John Micheal Jack Bob Jim Brian Chris Dave Mike

# 5 Applications

we have restricted our experiments to English, but in principle the techniques could apply to other languages; we are particularly interested in helping the emerging nations of Eastern Europe (including former Soviet Union) to develop linguistic technologies for 'new' national languages. These languages, e.g. Ukranian, have no Machine Tractable Dictionaries of computational grammars akin to, for example, LDOCE and ANLT for British English; to create these would require much effort by expert native-speaker linguists, so techniques for learning classifications and language models from a training Corpus are very attractive. Note that word-classification learning alone will not supply a compostional model of syntax and semantics of the sort used in many natural language processing or understanding systems (e.g. ANLT); but 'learnt' wordclasses *will* be useful in non-compositional language models as used in speech and handwriting recognition.

A common linguistic constraint model for speech and handwriting recognition is the $n$-pos or Markov model of word-tags. If a word-class clustering is computed using immediate neighbour collocation counts as the contextual distribution criterion, then a word-tagset will be derived which is purely Markovian. [Atwell 87] suggested that such a "pure Markovian" tagset should perform better in $n$-pos syntactic constraint models than linguistically-derived tagsets such as LOB or Brown for English; and certainly better than the absence of any tagset and tagged corpus for, say, Ukranian.

It remains to be shown how much of the structure of language can be uncovered with empiricist techniques; however, inference of word-classification is a useful first step towards the possibility posed by [Chomsky 57] of a "discovery procedure for grammars".

# References


**Eric Atwell.** *A Parsing Expert System which Learns from Corpus Analysis.* In W. Meijs (editor) - Corpus Linguistics and Beyond. Rodopi, Amsterdam. 1987

**Noam Chomsky.** *Syntactic Structures.* Mouton, The Hague. 1957

**Steven Finch.** *Finding Structure in Language.* PhD Thesis. Department of Cognitive Studies, Edinburgh University. 1993.

**Steve Hanlon and Roger Boyle.** *Syntactic Knowledge in Word Level Text Recognition.* In R. Beale and J. Finlay (editors) - Neural Networks and Pattern Recognition in HCI. Ellis Horwood. 1992.

**John Hughes.** *Automatically Acquiring a Classification of Words.* PhD Thesis. School of Computer Studies, University of Leeds. 1994.

**John Hughes and Eric Atwell.** *Acquiring and Evaluating a Classification of Words.* Proceedings of IEE Grammatical Inference Colloquium. University of Essex, Colchester. 1993.

**John Hughes and Eric Atwell.** *The Automated Evaluation of Inferred Word Classifications.* Proceedings of 11th European Conference on Artificial Intelligence. Amsterdam, The Netherlands. 1994.

**David Powers.** *On the Significance of Closed Classes and Boundary Conditions: Experiments in Lexical and Syntactic Learning.* In W. Daelemans and D.M.W. Powers (editors) - Background and Experiments in Machine Learning of Natural Language. Tilburg University. Institute for Language Technology and AI. pp. 245-266. 1992.

**Hinrich Schütze.** *Part-of-Speech Induction from Scratch.* Technical Report. Centre for the Study of Language and Information. Stanford. 1993.

**Jure Zupan.** *Clustering of Large Data Sets.* John Wiley and Sons, Chichester. 1982.