Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies

John G. McMahon*
The Queen's University of Belfast

Francis J. Smith

An automatic word-classification system has been designed that uses word unigram and bigram frequency statistics to implement a binary top-down form of word clustering and employs an average class mutual information metric. Words are represented as structural tags—*n-bit numbers the most significant bit-patterns of which incorporate class information. The classification system has revealed some of the lexical structure of English, as well as some phonemic and semantic structure. The system has been compared—directly and indirectly—with other recent word-classification systems. We see our classification as a means towards the end of constructing multilevel class-based interpolated language models. We have built some of these models and carried out experiments that show a 7% drop in test set perplexity compared to a standard interpolated trigram language model.*

1. Introduction

Many applications that process natural language can be enhanced by incorporating information about the probabilities of word strings; that is, by using statistical language model information (Church et al. 1991; Church and Mercer 1993; Gale, Church, and Yarowsky 1992; Liddy and Paik 1992). For example, speech recognition systems often require some model of the prior likelihood of a given utterance (Jelinek 1976). For convenience, the quality of these components can be measured by test set perplexity, *PP* (Bahl, Jelinek, and Mercer 1983; Bahl et al. 1989; Jelinek, Mercer, and Roukos 1990), in spite of some limitations (Ueberla 1994): $PP = \hat{P}(w_1^N)^{-\frac{1}{N}}$, where there are N words in the word stream $\langle w_1^N \rangle$ and $\hat{P}$ is some estimate of the probability of that word stream. Perplexity is related to entropy, so our goal is to find models that estimate a low perplexity for some unseen representative sample of the language being modeled. Also, since entropy provides a lower bound on the average code length, the project of statistical language modeling makes some connections with text compression—good compression algorithms correspond to good models of the source that generated the text in the first place. With an arbitrarily chosen standard test set, statistical language models can be compared (Brown, Della Pietra, Mercer, Della Pietra, and Lai 1992). This allows researchers to make incremental improvements to the models (Kuhn and Mori 1990). It is in this context that we investigate automatic word classification; also, some cognitive scientists are interested in those features of automatic word classification that have implications for language acquisition (Elman 1990; Redington, Chater, and Finch 1994).

One common model of language calculates the probability of the *i*th word w_i

* Department of Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland. E-mail: {J.McMahon,FJ.Smith}@qub.ac.uk

in a test set by considering the $n-1$ most recent words $\langle w_{i-n+1}, w_{i-n}, \ldots, w_{i-1}\rangle$, or $\langle w_{i-n+1}^{i-1}\rangle$ in a more compact notation. The model is finitary (according to the Chomsky hierarchy) and linguistically naive, but it has the advantage of being easy to construct and its structure allows the application of Markov model theory (Rabiner and Juang 1986).

Much work has been carried out on word-based n-gram models, although there are recognized weaknesses in the paradigm. One such problem concerns the way that n-grams partition the space of possible word contexts. In estimating the probability of the ith word in a word stream, the model considers all previous word contexts to be identical if and only if they share the same final $n-1$ words. This simultaneously fails to differentiate some linguistically important contexts and unnecessarily fractures others. For example, if we restrict our consideration to the two previous words in a stream—that is, to the trigram conditional probability estimate $\hat{P}(w_i|w_{i-2}^{i-1})$—then the sentences:

(1) a. The boys eat the sandwiches quickly.

and

(2) a. The cheese in the sandwiches is delicious.

contain points where the context is inaccurately considered identical. We can illustrate the danger of conflating the two sentence contexts by considering the nonsentences:

(1) b. *The boys eat the sandwiches is delicious.

and

(2) b. *The cheese in the sandwiches quickly.

There are some techniques to alleviate this problem—for example O'Boyle's n-gram ($n > 3$) weighted average language model (O'Boyle, Owens, and Smith 1994). A second weakness of word-based language models is their unnecessary fragmentation of contexts—the familiar sparse data problem. This is a main motivation for the multilevel class-based language models we shall introduce later. Successful approaches aimed at trying to overcome the sparse data limitation include backoff (Katz 1987), Turing-Good variants (Good 1953; Church and Gale 1991), interpolation (Jelinek 1985), deleted estimation (Jelinek 1985; Church and Gale 1991), similarity-based models (Dagan, Pereira, and Lee 1994; Essen and Steinbiss 1992), POS-language models (Derouault and Merialdo 1986) and decision tree models (Bahl et al. 1989; Black, Garside, and Leech 1993; Magerman 1994). We present an approach to the sparse data problem that shares some features of the similarity-based approach, but uses a binary tree representation for words and combines models using interpolation.

Consider the word ⟨`boys`⟩ in (1a) above. We would like to structure our entire vocabulary around this word as a series of similarity layers. A linguistically significant layer around the word ⟨`boys`⟩ is one containing all plural nouns; deeper layers contain more semantic similarities.

If sentences (1a) and (2a) are converted to the word-class streams ⟨`determiner noun verb determiner noun adverb`⟩ and ⟨`determiner noun preposition determiner noun verb adjective`⟩ respectively, then bigram, trigram, and possibly even higher n-gram statistics may become available with greater reliability for use as context differentiators (although Sampson [1987] suggests that no amount of word-class n-grams may be sufficient to characterize natural language fully). Of course, this still fails to differentiate many contexts beyond the scope of n-grams; while n-gram models of language may never fully model long-distance linguistic phenomena, we argue that it is still useful to extend their scope.

In order to make these improvements, we need access to word-class information (POS information [Johansson et al. 1986; Black, Garside, and Leech 1993] or semantic information [Beckwith et al. 1991]), which is usually obtained in three main ways: Firstly, we can use corpora that have been manually tagged by linguistically informed experts (Derouault and Merialdo 1986). Secondly, we can construct automatic part-of-speech taggers and process untagged corpora (Kupiec 1992; Black, Garside, and Leech 1993); this method boasts a high degree of accuracy, although often the construction of the automatic tagger involves a bootstrapping process based on a core corpus which has been manually tagged (Church 1988). The third option is to derive a fully automatic word-classification system from untagged corpora. Some advantages of this last approach include its applicability to any natural language for which some corpus exists, independent of the degree of development of its grammar, and its parsimonious commitment to the machinery of modern linguistics. One disadvantage is that the classes derived usually allow no linguistically sensible summarizing label to be attached (Schütze [1995] is an exception). Much research has been carried out recently in this area (Hughes and Atwell 1994; Finch and Chater 1994; Redington, Chater, and Finch 1993; Brill et al. 1990; Kiss 1973; Pereira and Tishby 1992; Resnik 1993; Ney, Essen, and Kneser 1994; Matsukawa 1993). The next section contains a presentation of a top-down automatic word-classification algorithm.

2. Word Classification and Structural Tags

Most statistical language models making use of class information do so with a single layer of word classes—often at the level of common linguistic classes: nouns, verbs, etc. (Derouault and Merialdo 1986). In contrast, we present the structural tag representation, where the symbol representing the word simultaneously represents the classification of that word (McMahon and Smith [1994] make connections between this and other representations; Black et al. [1993] contains the same idea applied to the field of probabilistic parsing; also structural tags can be considered a subclass of the more general tree-based statistical language model of Bahl et al. [1989]). In our model, each word is represented by an s-bit number the most significant bits of which correspond to various levels of classification; so given some word represented as structural tag w, we can gain immediate access to all s levels of classification of that word.

Generally, the broader the classification granularity we chose, the more confident we can be about the distribution of classes at that level, but the less information this distribution offers us about next-word prediction. This should be useful for dealing with the range of frequencies of n-grams in a statistical language model. Some n-grams occur very frequently, so word-based probability estimates can be used. However, as n-grams become less frequent, we would prefer to sacrifice predictive specificity for reliability. Ordinary POS-language models offer a two-level version of this ideal; it would be preferable if we could *defocus* our predictive machinery to some stages between all-word n-grams and POS n-grams when, for example, an n-gram distribution is not quite representative enough to rely on all-word n-grams but contains predictively significant divisions that would be lost at the relatively coarse POS level. Also, for rare n-grams, even POS distributions succumb to the sparse data problem (Sampson

1987); if very broad classification information was available to the language-modeling system, coarse-grained predictions could be factored in, which might improve the overall performance of the system in just those circumstances.

In many word-classification systems, the hierarchy is not explicitly represented and further processing, often by standard statistical clustering techniques, is required; see, for example, Elman (1990), Schütze (1993), Brill et al. (1990), Finch and Chater (1994), Hughes and Atwell (1994), and Pereira and Tishby (1992). With the structural tag representation, each tag contains explicitly represented classification information; the position of that word in class-space can be obtained without reference to the positions of other words. Many levels of classification granularity can be made available simultaneously, and the weight which each of these levels can be given in, for example, a statistical language model, can alter dynamically. Using the structural tag representation, the computational overheads for using class information can be kept to a minimum. Furthermore, it is possible to organize an n-gram frequency database so that close structural tags are stored near to each other; this could be exploited to reduce the search space explored in speech recognition systems. For example, if the system is searching for the frequency of a particular noun in an attempt to find the most likely next word, then alternative words should already be nearby in the n-gram database. Finally, we note that in the current implementation of the structural tag representation we allow only one tag per orthographic word-form; although many of the current word-classification systems do the same, we would prefer a structural tag implementation that models the multimodal nature of some words more successfully. For example, ⟨light⟩ can occur as a verb and as a noun, whereas our classification system currently forces it to reside in a single location.

Consider sentences (1a) and (2a) again; we would like to construct a clustering algorithm that assigns some unique s-bit number to each word in our vocabulary so that the words are distributed according to some approximation of the layering described above—that is, ⟨boys⟩ should be close to ⟨people⟩ and ⟨is⟩ should be close to ⟨eat⟩. We would also like semantically related words to cluster, so that, although ⟨boys⟩ may be near ⟨sandwiches⟩ because both are nouns, ⟨girls⟩ should be even closer to ⟨boys⟩ because both are human types. In theory, structural tag representations can be dynamically updated—for example, ⟨bank⟩ might be close to ⟨river⟩ in some contexts and closer to ⟨money⟩ in others. Although we could construct a useful set of structural tags manually (McMahon 1994), we prefer to design an algorithm that builds such a classification.

For a given vocabulary V, the mapping t initially translates words into their corresponding unique structural tags. This mapping is constructed by making random word-to-tag assignments.

The mutual information (Cover and Thomas 1991) between any two events x and y is:

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(x)}$$

If the two events x and y stand for the occurrence of certain word-class unigrams in a sample, say c_i and c_j, then we can estimate the mutual information between the two classes. In these experiments, we use maximum likelihood probability estimates based on a training corpus. In order to estimate the average class mutual information for a classification depth of s bits, we compute the average class mutual information:

$$M_s(t) = \sum_{c_i,c_j} P(c_i,c_j) \times \log \frac{P(c_i,c_j)}{P(c_i)P(c_j)} \qquad (1)$$

where c_i and c_j are word classes and $M_s(t)$ is the average class mutual information for structural tag classification t at bit depth s. This criterion is the one used by Brown, Della Pietra, DeSouza, Lai, and Mercer (1992); Kneser and Ney (1993) show how it is equivalent to maximizing the bi-POS-language model probability. We are interested in that classification which maximizes the average class mutual information; we call this t^o and it is found by computing:

$$M_s(t^o) = \max_t M_s(t) \tag{2}$$

Currently, no method exists that can find the globally optimal classification, but suboptimal strategies exist that lead to useful classifications. The suboptimal strategy used in the current automatic word-classification system involves selecting the *locally* optimal structure between t and t', which differ only in their classification of a single word. An initial structure is built by using the computer's pseudorandom number generator to produce a random word hierarchy. Its $M(t)$ value is calculated. Next, another structure, t' is created as a copy of the main one, with a single word moved to a different place in the classification space. Its $M(t')$ value is calculated. This second calculation is repeated for each word in the vocabulary and we keep a record of the transformation which leads to the highest $M(t')$. After an iteration through the vocabulary, we select that t' having the highest $M(t')$ value and continue until no single move leads to a better classification. With this method, words which at one time are moved to a new region in the classification hierarchy can move back at a later time, if licensed by the mutual information metric. In practice, this does happen. Therefore, each transformation performed by the algorithm is not irreversible within a level, which should allow the algorithm to explore a larger space of possible word classifications.

The algorithm is embedded in a system that calculates the best classifications for all levels beginning with the highest classification level. Since the structural tag representation is binary, this first level seeks to find the best distribution of words into two classes. Other versions of the top-down approach are used by Pereira and Tishby (1992) and Kneser and Ney (1993) to classify words; top-down procedures are also used in other areas (Kirkpatrick, Gelatt, and Vecchi 1983). The system of Pereira and Tishby (1992; Pereira, Tishby, and Lee 1993) has the added advantage that class membership is probabilistic rather than fixed.

When the locally optimal two-class hierarchy has been discovered by maximizing $M_1(t)$, whatever later reclassifications occur at finer levels of granularity, words will always remain in the level 1 class to which they now belong. For example, if many nouns now belong to class 0 and many verbs to class 1, later subclassifications will not influence the $M_1(t)$ value. This reasoning also applies to all classes $s = 2, 3 \ldots 16$ (see Figure 1).

We note that, in contrast with a bottom-up approach, a top-down system makes its first decisions about class structure at the root of the hierarchy; this constrains the kinds of classification that may be made at lower levels, but the first clustering decisions made are based on healthy class frequencies; only later do we start noticing the effects of the sparse data problem. We therefore expect the topmost classifications to be less constrained, and hopefully more accurate. With a bottom-up approach, the reverse may be the case. The tree representation also imposes its own constraints, mentioned later.

This algorithm, which is $O(V^3)$ for vocabulary size V, works well with the most

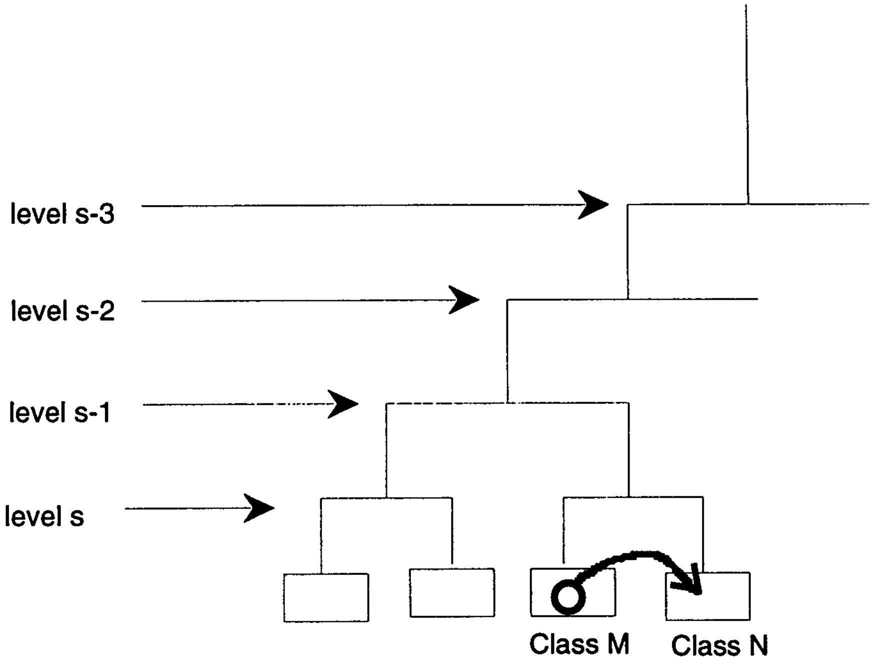


Figure 1
Top-down clustering. The algorithm is designed so that, at a given level s, words will have already been re-arranged at levels $s-1$, etc. to maximize the average class mutual information. Any alterations at level s will not bear on the classification achieved at $s-1$. Therefore, a word in class M may only move to class N to maximize the mutual information—any other move would violate a previous level's classification.

frequent words from a corpus[1]; however, we have developed a second algorithm, to be used after the first, to allow vocabulary coverage in the range of tens of thousands of word types. This second algorithm exploits Zipf's law (1949)—the most frequent words account for the majority of word tokens—by adding in low-frequency words only after the first algorithm has finished processing high-frequency ones. We make the assumption that any influence that these infrequent words have on the first set of frequent words can be discounted. The algorithm is an order of magnitude less computationally intensive and so can process many more words in a given time. By this method, we can also avoid modeling only a simplified subset of the phenomena in which we are interested and hence avoid the danger of designing systems that do not scale-up adequately (Elman 1990). Once the positions of high-frequency words has been fixed by the first algorithm, they are not changed again; we can add any new word, in order of frequency, to the growing classification structure by making 16 binary decisions: Should its first bit be a 0 or a 1? And its second? Of our 33,360 word vocabulary, we note that the most frequent 569 words are clustered using the main

[1] In a worst case analysis, the mutual information metric will be $O(V^2)$ and we need to evaluate the tree on V occasions—each time with one word reclassified; lower order terms (for example, the number of iterations at each level) can be ignored. In practice, the mutual information calculation is much less than $O(V^2)$ since there are far fewer than V^2 bigrams observed in our training text.

algorithm; the next 15,000 are clustered by our auxiliary algorithm and the remaining 17,791 words are added to the tree randomly. We add these words randomly due to hardware limitations, though we notice that the 15,000th most frequent word in our vocabulary occurs twice only—a very difficult task for any classification system. The main algorithm takes several weeks to cluster the most frequent 569 words on a Sparc-IPC and several days for the supplementary algorithm.

3. Word Classification Performance

In evaluating this clustering algorithm, we were interested to see if it could discover some rudiments of the structures of language at the phonemic, syntactic, and semantic levels; we also wanted to investigate the possibility that the algorithm was particularly suited to English.

To these ends, we applied our algorithm to several corpora. It successfully discovered major noun-verb distinctions in a toy regular grammar introduced by Elman (1990), made near perfect vowel-consonant distinctions when applied to a phonemic corpus and made syntactic and semantic distinctions in a Latin corpus (McMahon 1994). It also discovered some fine-grained semantic detail in a hybrid POS-word corpus. However, classification groups tended to be dispersed at lower levels; we shall discuss this phenomenon with respect to the distribution of number words and offer some reasons in a later section.

3.1 Clustering Results

We report on the performance of our top-down algorithm when applied to the most frequent words from an untagged version of the LOB corpus (Johansson et al. 1986) and also when applied to a hybrid word-and-class version of the LOB. We used structural tags 16 bits long and we considered the 569 most frequent words; this gave us 46,393 bigrams to work with—all other word bigrams were ignored. We present the following figures as illustrations of the clustering results: our main use for the classification system will be as a way to improve statistical language models; we eschew any detailed discussion of the linguistic or cognitive relevance of the clustering results. Illustrative clusterings of this type can also be found in Pereira, Tishby, and Lee (1993), Brown, Della Pietra, Mercer, Della Pietra, and Lai (1992), Kneser and Ney (1993), and Brill et al. (1990), among others.

In Figure 2, we observe the final state of the classification, to a depth of five bits. Many syntactic and some semantic divisions are apparent—prepositions, pronouns, verbs, nouns, and determiners cluster—but many more distinctions are revealed when we examine lower levels of the classification. For example, Figure 3 shows the subcluster of determiners whose initial structural tag is identified by the four-bit schema 0000. In Figure 4 we examine the finer detail of a cluster of nouns. Here, some semantic differences become clear (we have internally ordered the words to make the semantic relations easier to spot). Many of the 2^5 groups listed in Figure 2 show this type of fine detail. It is clear, also, that there are many anomalous classifications, from a linguistic point of view. We shall say more about this later.

In a second experiment (see Figure 5), a hybrid version of the LOB corpus was created: we replaced each word and part-of-speech pair by the word alone if the part-of-speech was a singular noun, the base form of a verb, or the third person singular present tense of a verb; otherwise we replaced it by the part-of-speech. By doing this, we hoped to lighten the burden of inducing syntactic structure in the vocabulary to see if the classification system could move beyond syntax and into semantic clustering. We considered that, of the word tokens replaced by their part-

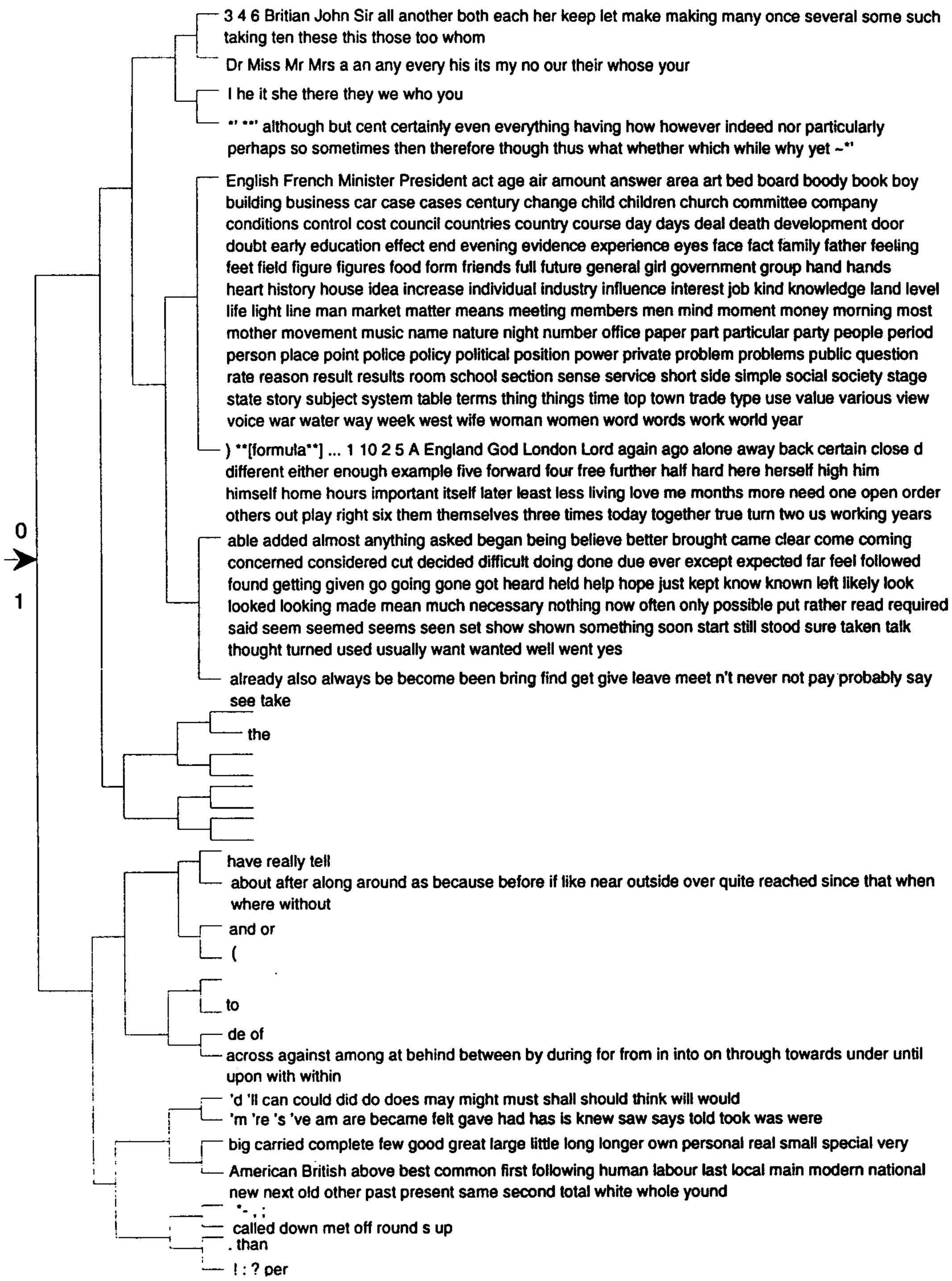

3 4 6 Britian John Sir all another both each her keep let make making many once several some such taking ten these this those too whom
Dr Miss Mr Mrs a an any every his its my no our their whose your
I he it she there they we who you
" "" although but cent certainly even everything having how however indeed nor particularly perhaps so sometimes then therefore though thus what whether which while why yet -"
English French Minister President act age air amount answer area art bed board boody book boy building business car case cases century change child children church committee company conditions control cost council countries country course day days deal death development door doubt early education effect end evening evidence experience eyes face fact family father feeling feet field figure figures food form friends full future general girl government group hand hands heart history house idea increase individual industry influence interest job kind knowledge land level life light line man market matter means meeting members men mind moment money morning most mother movement music name nature night number office paper part particular party people period person place point police policy political position power private problem problems public question rate reason result results room school section sense service short side simple social society stage state story subject system table terms thing things time top town trade type use value various view voice war water way week west wife woman women word words work world year
) **[formula**] ... 1 10 2 5 A England God London Lord again ago alone away back certain close d different either enough example five forward four free further half hard here herself high him himself home hours important itself later least less living love me months more need one open order others out play right six them themselves three times today together true turn two us working years
0
1
able added almost anything asked began being believe better brought came clear come coming concerned considered cut decided difficult doing done due ever except expected far feel followed found getting given go going gone got heard held help hope just kept know known left likely look looked looking made mean much necessary nothing now often only possible put rather read required said seem seemed seems seen set show shown something soon start still stood sure taken talk thought turned used usually want wanted well went yes
already also always be become been bring find get give leave meet n't never not pay probably say see take
the
have really tell
about after along around as because before if like near outside over quite reached since that when where without
and or
(
to
de of
across against among at behind between by during for from in into on through towards under until upon with within
'd 'll can could did do does may might must shall should think will would
'm 're 's 've am are became felt gave had has is knew saw says told took was were
big carried complete few good great large little long longer own personal real small special very
American British above best common first following human labour last local main modern national new next old other past present same second total white whole yound
*- , ;
called down met off round s up
. than
! : ? per

Figure 2
Final distribution of the most frequent words from the LOB corpus. Only the first five levels of classification are given here, but important syntactic relations are already clear. The empty classes are shown to make the binary topology clear.

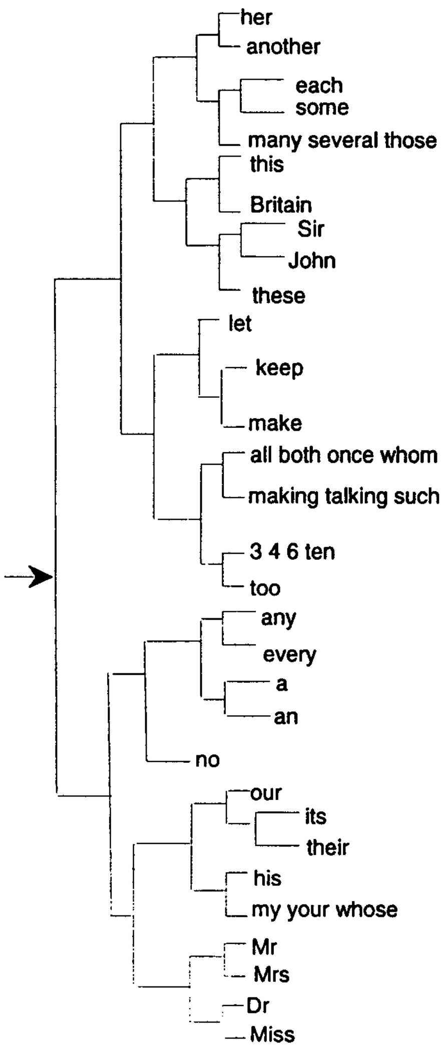


Figure 3
Word classification results. Detail of relationship between words whose final tag value starts with the four bits 0000. Many of these words exhibit determiner-like behavior.

of-speech, the vast majority would be function words and hence would contribute little to any semantic classification. Also, we hoped that relatively rare content words would now find themselves within slightly more substantial bigram contexts, again aiding the clustering process.

When we examine the most frequent "words" of this hybrid corpus, we find that there are many more content words present, but that the remaining content words still have an indirect effect on word classification, since they are represented by the part-of-speech of which they are an example. Figure 5 shows many of the largest groupings of words found after processing, at a classification level of nine bits. By inspection, we observe a variety of semantic associations, although there are many obvious anomalies. We offer several explanations for this—words are polysemic in reality, the clustering criterion exploits bigram information only, and this algorithm,

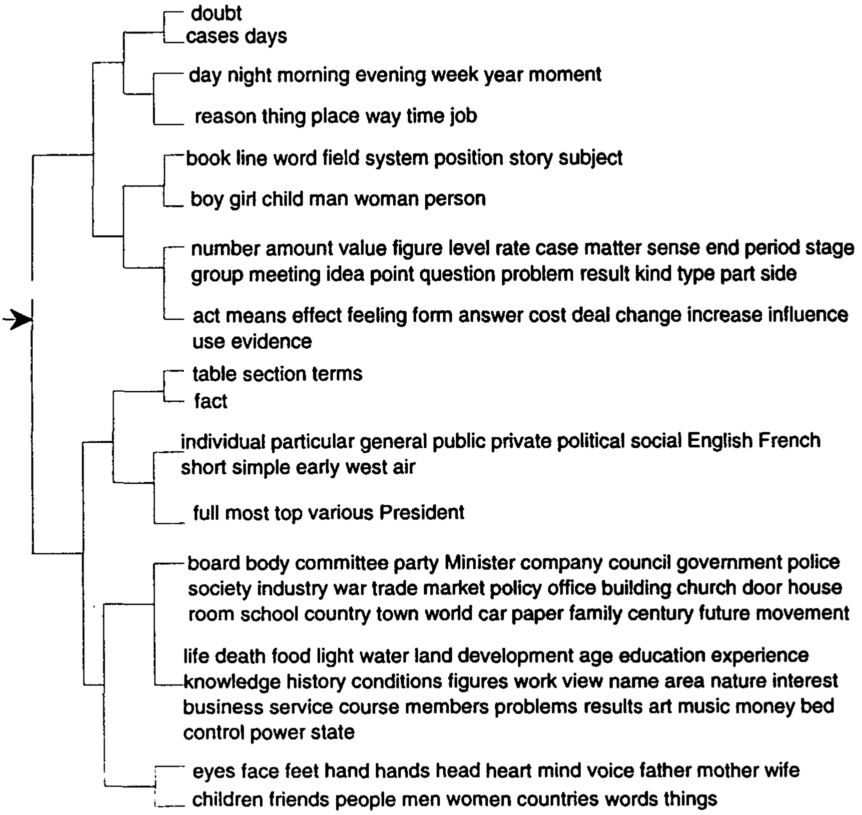

doubt
cases days
day night morning evening week year moment
reason thing place way time job
book line word field system position story subject
boy girl child man woman person
number amount value figure level rate case matter sense end period stage group meeting idea point question problem result kind type part side
act means effect feeling form answer cost deal change increase influence use evidence
table section terms
fact
individual particular general public private political social English French
short simple early west air
full most top various President
board body committee party Minister company council government police society industry war trade market policy office building church door house room school country town world car paper family century future movement
life death food light water land development age education experience knowledge history conditions figures work view name area nature interest business service course members problems results art music money bed control power state
eyes face feet hand hands head heart mind voice father mother wife
children friends people men women countries words things

Figure 4
Detail, from Level 5 to Level 9, of many noun-like words. Semantic differences are registered.

like others, finds a locally optimal classification. In each word group we include here, the entire membership is listed (except for the irrelevant POS tags). The remaining groups not presented here also display strong semantic clustering. From this second experiment, we conclude that bigram statistics can be used to make some semantic differentiations. This lends support to our case for using multilevel statistical language models—we can see the kinds of distinctions that structural tags can make and which will be lost in the more usual two-level (word and POS) language models.

Finally, Figure 6 shows the complete phoneme classification of a phonemic version of the VODIS corpus. The most obvious feature of this figure is the successful distinction between vowels and consonants. Beyond the vowel-consonant distinction, other similarities emerge: vowel sounds with similar vocal tract positions are clustered closely—the /a/ sounds, for example, and the phoneme pair /o/ and /oo/; some consonants that are similarly articulated also map onto local regions of the classification space—/r/ and /rx/, /ch/ and /z/, and /n/ and /ng/, for example.

3.2 Clustering Comparisons

A scientifically reliable method of comparing classifications would be to measure how different they are from randomly generated classifications. This kind of approach has been taken by Redington, Chater, and Finch (1994) but is not used here because it is apparent that the classifications are clearly different from random ones and, more significantly, because many classification processes could produce distributions that arm breath breathing cheek chin coat eye fist forehead hair handkerchief hat head knee lad memory mouth neck nose purse shoulder skirt throat whistle wrist ability curiosity destination discretion employer hairdresser heritage inheritance intention tone irritation lifetime loneliness midst mistress profession reluctance typewriter career castle

aunt brother father father-in-law husband mother sister uncle wife gaze jaw lip voice diary wallet companion lordship partner wholesaler

approval charm conscience contempt disposal embarrassment engagement fate fault good ideal imagination mind name opinion resignation sake soul speech temper thesis vision will arrival audience correspondent tutor designer friend neighbour lord lover environment childhood cousin daughter son niece elbow hand tongue lap apartment cave surgery uniform

decade year month fortnight week hour inch lot spot step string host row pool pot ending matter instant minute moment while pause second bit minimum way clue chance fool illusion joke gasp mistake nuisance offence pity reason proposition enquiry job ship compartment room chair leg ridge blanket altar envelope powder adult debt determination disposition

actor actress artist boy bride captain catholic chap child citizen composer couple critic doctor engineer fellow gentleman girl god hostess individual journalist king lady lawyer legend man novelist observer painter patient people person priest prisoner producer psychologist queen ruler scholar scientist singer soldier sovereign stranger teacher widow woman writer bird cow creature dog mantis rat tree assembly republic accident incident situation experiment target corruption scandal struggle armistice invitation obligation opportunity temptation tendency willingness desire passion mood fortune multitude maximum alternative amendment proposal decision document lease catechism inscription phrase word poem attempt effort gesture sigh whisper message bell knife pen meal poultice clock taxi cloud lawn staircase device explosive explosion shot ray shower dress cathedral hut

living-room lounge attic roof bathroom bedroom dining-room drawing reception drawing-room kitchen sitting-room carpet floor window cooker oven cabinet desk lamp mirror cupboard shelf telephone sofa hearth fire fireplace flame turf corner corridor hall door doorway entrance tunnel lock key gate background rear boundary hedge wall yard ground surface square garage garden terrace farm farmhouse outside field courtyard barn stable cabin cage pit cell cottage flat studio hotel house laboratory palace tower bay beach lake pond river pier mud net city town village neighbourhood country nation landscape park forest hill slope mine lane road street island moon moonlight sun sky storm wind weather horse tractor continent world landlord owner maid manufacturer peasant cafe clinic concert funeral inquest hearing honeymoon wedding airport platform helicopter boat bus car coach plane flight journey mission model bishop vicar parish church pulpit throne title temple devil gospel truth oath trio tribe budget loan pension penalty prize exchequer taxpayer campaign presentation election term conservative ambulance wound brain baby body finger chest waist stomach womb heart flesh skin heel collar pocket bomb receiver tape gun pistol rifle bullet sergeant drillbriefcase dictionary magazine camera cigarette jacket bow shaft stem sword wing ontrary craft crying dark week-end weekend evening future past defendant registrar spectator student subject editor grandfather downwash flood trawl incidence left reverse outset peak waltz wolf nucleus

feels regards sees thinks knows likes loves wants wishes seeks calls asks changes describes says uses plays owns loses

admit assume believe think understand realise realize reckon conclude remember decide dare deserve expect suppose seem tend appear prefer want own bother afford refuse fail feel forget hate hesitate intend know learn like begin propose try continue cease

acts arises begins consists depends lies occurs refers rests varies

assumption certainty doubt hope wonder

endeavour excuse hurry need request risk urge right sign

Figure 5
Semantic clustering results. Memberships of 11 of the 38 classes whose size is greater than ten, at a classification level of nine bits. From the LOB corpus. Body parts, relatives, mental states, human roles, house parts, two classes of mental verb, relation verbs, hope-nouns, effort verbs. The training corpus was the hybrid POS-word text.

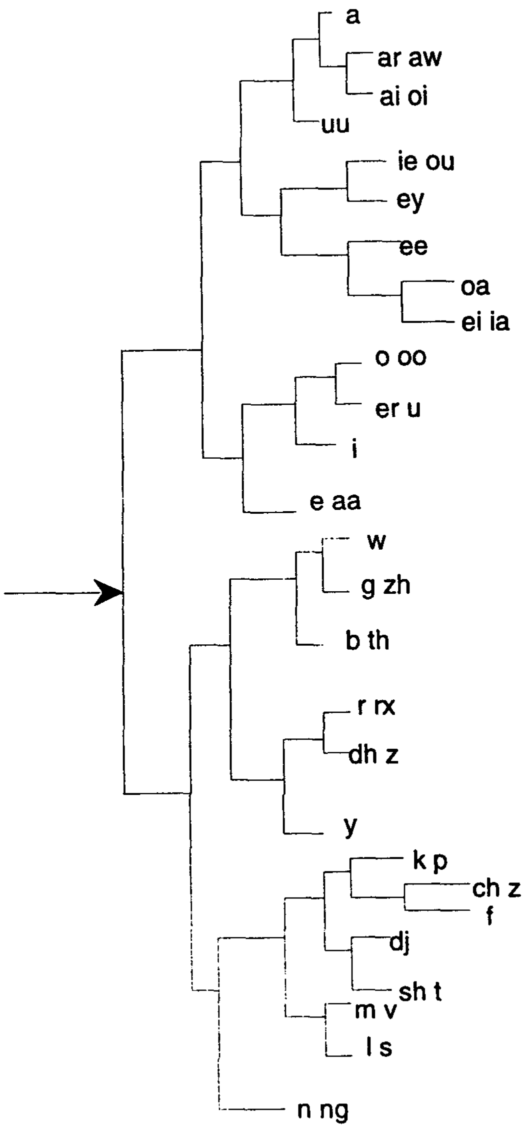


Figure 6
Automatic phoneme clustering. Differentiation between vowels and consonants. From a phonemic version of the VODIS corpus.

are nonrandom, but have nothing to do with lexical categories and are not useful for class-based statistical language modeling.

The question of the criterion of a successful classification is dependent upon research motivations, which fall into two broad schools. The first school is made up of those who primarily would like to recover the structures linguists posit—the structures they seek are mainly syntactic but can also be semantic. The second school is interested in classifications that help to improve some language model or other language-processing system and that may or may not exhibit linguistically perspicuous categories. Unless modern linguistics is radically wrong, a degree of overlap should occur in these two ideals.

Researchers who claim that linguistically well-formed classifications are not the immediate goal of their research must find some other way of measuring the applicability of their classifications. For example, we can operationally define good word classifications as those conferring performance improvements to statistical language models. We shall make this our goal later, but first we compare our system with others by inspection.

In Brill et al. (1990), another automatic word-classification algorithm was developed and trained using the Brown corpus; they report success at partitioning words into word classes. They note that pronouns have a disjoint classification, since the `+nominative` and `-nominative` pronouns—for example, ⟨`I`⟩, ⟨`they`⟩ and ⟨`me`⟩, ⟨`them`⟩ respectively—have dissimilar distributions. These effects are replicated in our experiment. They report other, more fine-grained features such as possessives, singular determiners, definite-determiners and *Wh*-adjuncts. Our algorithm also distinguishes these features. Brill et al. do not report any substantial adjective clustering, or noun clustering, or singular-plural differences, or co-ordinating and subordinating conjunction distinction, or verb tense differentiation. At lower levels, the only semantic clustering they report involves the group: ⟨`man world time life work people years`⟩ and the group: ⟨`give make take find`⟩.

The results described in Brown, Della Pietra, DeSouza, Lai, and Mercer (1992) are based on a training set two orders of magnitude greater than the one used in the above experiment. Even the vocabulary size is an order of magnitude bigger. As the vocabulary size is increased, the new vocabulary items tend, with a probability approaching unity, to be content words: after approximately one thousand words, few function words are left undiscovered. This increase in resources makes contexts more balanced and, simultaneously, more statistically significant. It also allows many more content words to be grouped together semantically. The authors give two tables of generated word classes, one being specially selected by them and the other containing randomly selected classes. They do not report on any overall taxonomic relations between these classes, so it is not possible to compare the broad detail of the two sets of data.

The results of Finch and Chater (1992, 1991) are also based on a substantially larger corpus. Finch and Chater also run a version of the Elman experiment (see below). Their system fails to produce a complete noun-verb distinction at the highest level, though they offer an argument to suggest that the inadequacy lies in the nature of Elman's pseudo–natural language corpus; our system uses Elman's corpus but succeeds in making the primary noun-verb distinction. Finch and Chater also cluster letters and phonemes—their system succeeds in distinguishing between vowels and consonants in the letter experiment, and only the phoneme /u/ is incorrectly classified in the phoneme experiment. Conversely, our algorithm completely clusters phonemes into vowels and consonants, but performs less well with letters (McMahon 1994).

Pereira and Tishby (1992) do not give details of syntactic similarity—they concentrate on a small number of words and make fine-grained semantic differentiations between them. Their evaluation techniques include measuring how helpful their system is in making selectional restrictions and in disambiguating verb-noun pairs.

Schütze (1993) uses a standard sparse matrix algorithm with neural networks; his system is the only one that attempts to side-step the problem of deciding what his clusters are clusters *of*, by producing a system that generates its own class labels. He does not report the overall structure of his one-level classification. His training set is one order of magnitude bigger than the largest one used in the present experiments.

3.2.1 Ceteris Paribus Qualitative Comparison. This section describes comparisons between our algorithm and others, where some of the experimental parameters are controlled—for example, corpus size. We considered it useful to compare the performance of our algorithm with others' on precisely the same input data because we believe that factors like vocabulary, corpus size, and corpus complexity make evaluation difficult.

A Recurrent Neural Network and a Regular Grammar. We redescribe the salient details of one of the experiments performed by Elman (1990). The grammar that generates the language upon which this experiment is based is, according to the Chomsky classification, type 4 (regular, or finite-state). Its production rules are shown in Figure 7. Some of the words belong to two or more word classes. The sentence frames encode a simple semantics—noun types of certain classes engage in behavior unique to that class. Elman generates a 10,000-sentence corpus to be used as the training corpus. Each sentence frame is just as likely to be selected as any other; similarly, each word member of a particular word group has an equiprobable chance of selection. No punctuation is included in the corpus, so sentence endings are only implicitly represented—for example, the segment stream ⟨`cat smell cookie dog exist boy smash plate`⟩ contains a three-word sentence followed by a two-word sentence followed by another three-word sentence.

After training, Elman's net was tested on an unseen set, generated by the same underlying grammar. The network's performance was poor—only achieving a prediction error rate slightly above chance. Elman then presented the training data to the net a further four times, but the prediction was still poor. He claimed that, with even more training, the net could have improved its performance further. But this was not the main goal of the experiment; instead, hierarchical cluster analysis was performed on the averaged hidden unit activations for each of the 29 words. Figure 8 reproduces the similarity tree that cluster analysis of the recurrent net produced. His analysis reveals that the network has learned most of the major syntactic differences and many of the semantic ones coded in the original language. For example, there is a clear distinction between verbs and nouns; within the main class of nouns, there is a clear animate-inanimate distinction; within that, the classes of agent-patient, aggressor, and nonhuman animal have been induced. The analysis is not perfect: the most important distinction is considered to be between a handful of inanimate noun objects (bread, cookie, sandwich, glass, and plate) and the rest of the vocabulary.

We now discuss the results obtained when our algorithm is applied to a similar test corpus. Elman's grammar of Figure 7 was used to produce a corpus of 10,000 sentences with no sentence breaks. Unigram and bigram word-frequency statistics were generated. Our structural tag word-classification algorithm was applied to the initial mapping, which randomly assigned tag values to the 29 words. Figure 9 shows the important classification decisions made by this algorithm. Unlike the Elman classification (see Figure 8), informationally useful class structure exists from level 1 onwards. This algorithm also produces a classification some features of which are qualitatively better than Elman's—all nouns and all verbs are separated; all animates and inanimates are separated. The multicontext noun/verb ⟨`break`⟩ is identified as different from other verbs; intransitive verbs cluster together and the aggressive nouns are identified. This algorithm does not recapture the complete syntax and semantics of the language—human nouns and non-aggressive animate nouns remain mixed, and the food noun cluster failed to attract the word ⟨`sandwich`⟩. This experiment was repeated several times, each time resulting in a classification whose overall structure was similar but whose fine detail was slightly different. One run, for example, correctly differentiated

S → NOUN-HUMAN VERB-EAT NOUN-FOOD

S → NOUN-HUMAN VERB-PERCEPT NOUN-INAN

S → NOUN-HUMAN VERB-DESTROY NOUN-FRAGILE

S → NOUN-HUMAN VERB-INTRAN

S → NOUN-HUMAN VERB-TRAN NOUN-HUMAN

S → NOUN-HUMAN VERB-AGPAT NOUN-INAN

S → NOUN-HUMAN VERB-AGPAT

S → NOUN-ANIM VERB-EAT NOUN-FOOD

S → NOUN-ANIM VERB-TRAN NOUN-ANIM

S → NOUN-ANIM VERB-AGPAT NOUN-INANIM

S → NOUN-ANIM VERB-AGPAT

S → NOUN-INAN VERB-AGPAT

S → NOUN-AGRESS VERB-DESTROY NOUN-FRAGILE

S → NOUN-AGRESS VERB-EAT NOUN-HUMAN

S → NOUN-AGRESS VERB-EAT NOUN-ANIM

S → NOUN-AGRESS VERB-EAT NOUN-FOOD

NOUN-HUMAN → man woman girl boy

NOUN-ANIM → cat mouse dog man woman girl boy dragon monster lion

NOUN-INAN → book rock car cookie break bread sandwich glass plate

NOUN-AGRESS → dragon monster lion

NOUN-FRAGILE → glass plate

NOUN-FOOD → cookie break bread sandwich

VERB-INTRAN → think sleep exist

VERB-TRAN → see chase like

VERB-AGPAT → move break

VERB-PERCEPT → smell see

VERB-DESTROY → break smash

VERB-EAT → eat

Figure 7
The Elman grammar. There are 16 nonterminal rules and 12 terminals. Notice also that terminals can belong to more than one word class—for example, ⟨break⟩ is an inanimate noun, a food noun, an agent-patient verb, and a destroy verb.

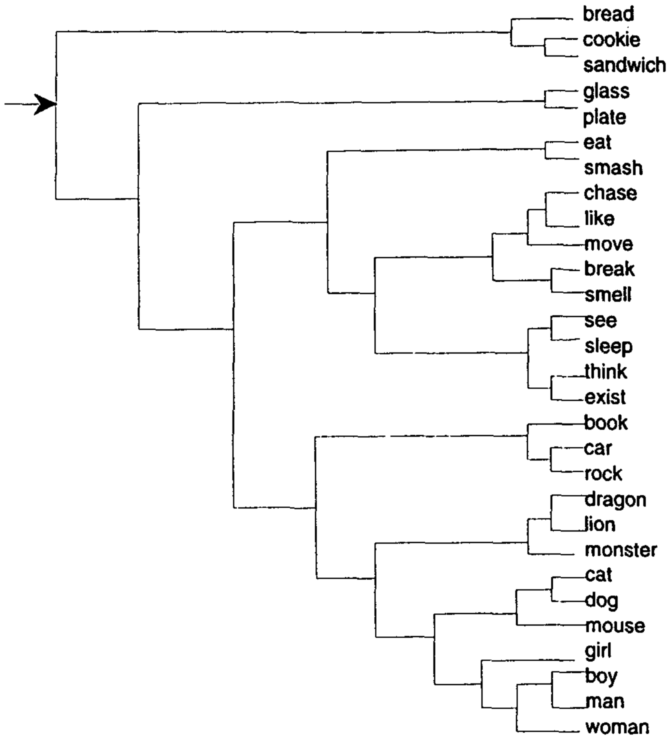


Figure 8
Elman's results. A Cluster analysis of the hidden units of a trained recurrent net, showing the major verb-noun distinction, as well as many other syntactic and semantic fine-grained distinctions.

between small animals and humans, but failed to recognize food nouns as a complete group. Another run identified food nouns perfectly but failed to separate aggressors from other animates.

Classification Using a Merging Algorithm. The systems described in Brown, Della Pietra, DeSouza, Lai, and Mercer (1992) and Brill and Marcus (1992) both provide examples of bottom-up, merge-based classification systems; a version of such a system was chosen to be implemented and tested against our algorithm, using the same input data. The Brown system uses a principle of class merging as its main clustering technique. The initial classification contains as many classes as there are words to classify, each word in its own class. Initially these classes are all mutually independent. Then two classes are chosen to merge; the criterion of choice is based on a mutual information calculation (see Equation 2). The process is repeated until only one class remains. Next, the *order* of merging provides enough information for a hierarchical cluster to be constructed. A comparison experiment was designed using the 70,000-word VODIS corpus (Cookson 1988) as a source of frequency information; our system and the merging system were given a set of those words from the decapitalized and depunctuated corpus (except for the apostrophe when it is a part of a word) whose frequencies were greater than 30. This accounted for the 256 most frequent words.

The final classifications, to a depth of five levels, are shown in Figure 10 and Figure 11 for the bottom-up and top-down systems, respectively. The difficulty of com-

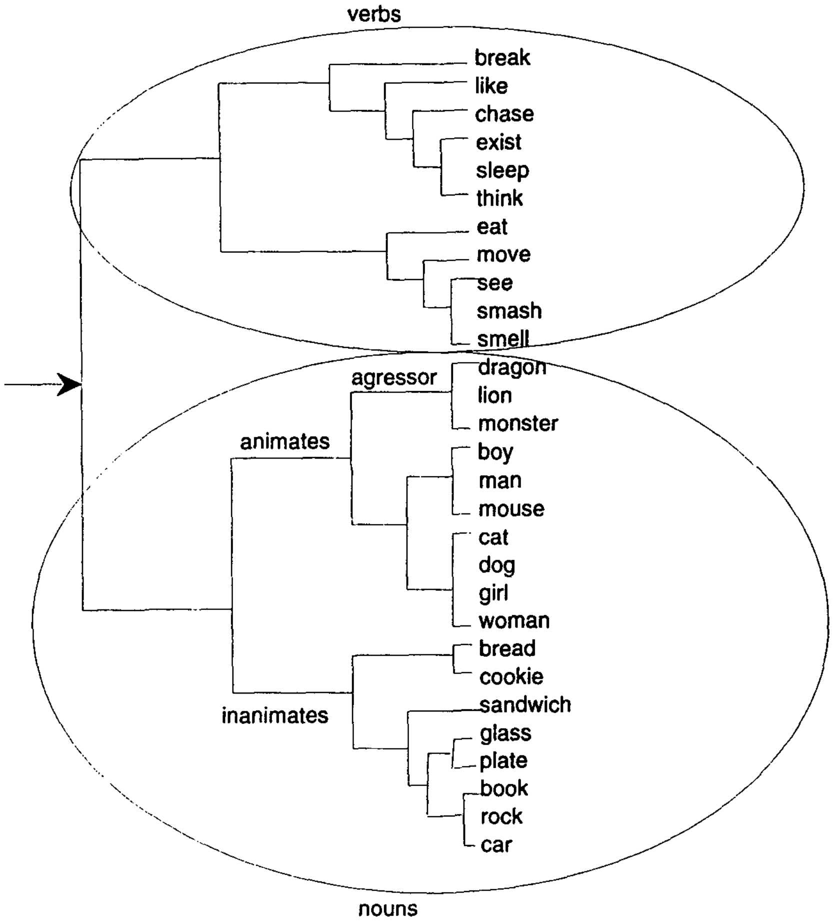


Figure 9
Class hierarchy using structural tags and average class mutual information for the Elman experiment. Subclassifications are only displayed up to the first point where they are misclassified or when they correctly identify a class.

parative evaluation becomes more apparent when looking at these two classification techniques; both seem to have strengths and weaknesses—for example, our system possesses a more balanced overall structure, compared to the merge-based system, which has as its most important distinction the difference between number words and all other words. It should be stated that both systems successfully discover a large degree of syntactic structure (e.g., prepositions, determiners, nouns, verbs, adverbs, and conjunctions) and some semantic structure (e.g., place names), given the relatively small size of the corpus.

With nouns, at the first branching of the trees the merge-based system has a 39/44 split compared to the top-down system, which has an 11/72 split. The merge-base system performs better when clustering the modals ⟨`can could do did should will would don't can't haven't`⟩—these are almost all still together at a depth of five bits, compared to two bits for the top-down system. The top-down system clusters the number words ⟨`two three four five six seven eight nine ten eleven twelve thirteen sixteen`⟩ to Level 5; to Level 4 the following numbers are also included: ⟨`fifteen seventeen fourty fifty sixty eighty`⟩; later we offer reasons

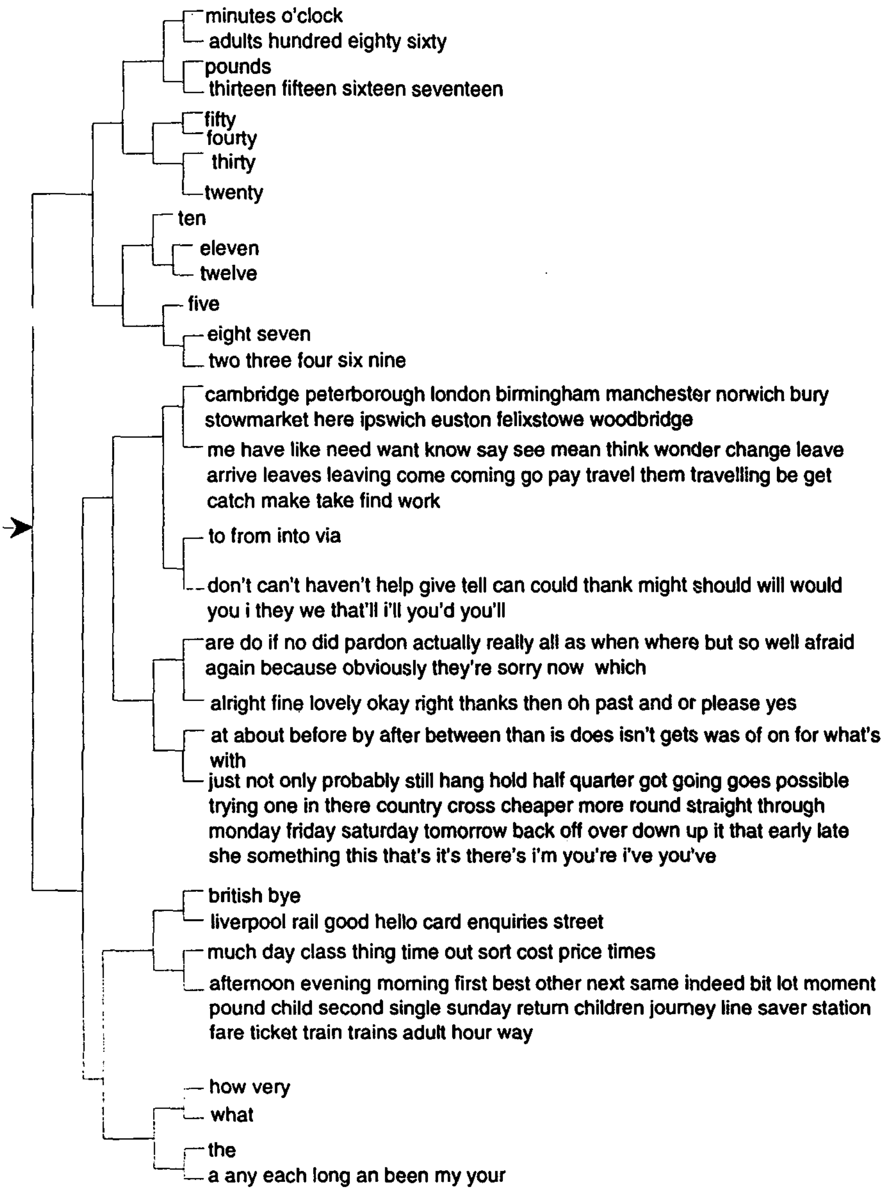

minutes o'clock
adults hundred eighty sixty
pounds
thirteen fifteen sixteen seventeen
fifty
fourty
thirty
twenty
ten
eleven
twelve
five
eight seven
two three four six nine
cambridge peterborough london birmingham manchester norwich bury stowmarket here ipswich euston felixstowe woodbridge
me have like need want know say see mean think wonder change leave arrive leaves leaving come coming go pay travel them travelling be get catch make take find work
to from into via
don't can't haven't help give tell can could thank might should will would you i they we that'll i'll you'd you'll
are do if no did pardon actually really all as when where but so well afraid again because obviously they're sorry now which
alright fine lovely okay right thanks then oh past and or please yes
at about before by after between than is does isn't gets was of on for what's with
just not only probably still hang hold half quarter got going goes possible trying one in there country cross cheaper more round straight through monday friday saturday tomorrow back off over down up it that early late she something this that's it's there's i'm you're i've you've
british bye
liverpool rail good hello card enquiries street
much day class thing time out sort cost price times
afternoon evening morning first best other next same indeed bit lot moment pound child second single sunday return children journey line saver station fare ticket train trains adult hour way
how very
what
the
a any each long an been my your

Figure 10
Merge-based clustering. Classification of the most frequent words of a formatted VODIS corpus, using a merge-based method.

for this kind of division between numbers. The merge-based system includes these numbers at Level 1 only. Both systems identify place names well (⟨`liverpool`⟩ is most often seen in this corpus as part of the place name—⟨`liverpool street station`⟩); however, the nearest class in the merge-based system is a group of verb-like words, whereas in the top-down system, the nearest class is a group of nouns. With verbs, the merge-based system performs better, producing a narrower dispersion of words. However, this success is slightly moderated by the consideration that one half of the entire merge classification system was allocated to number words, leaving the

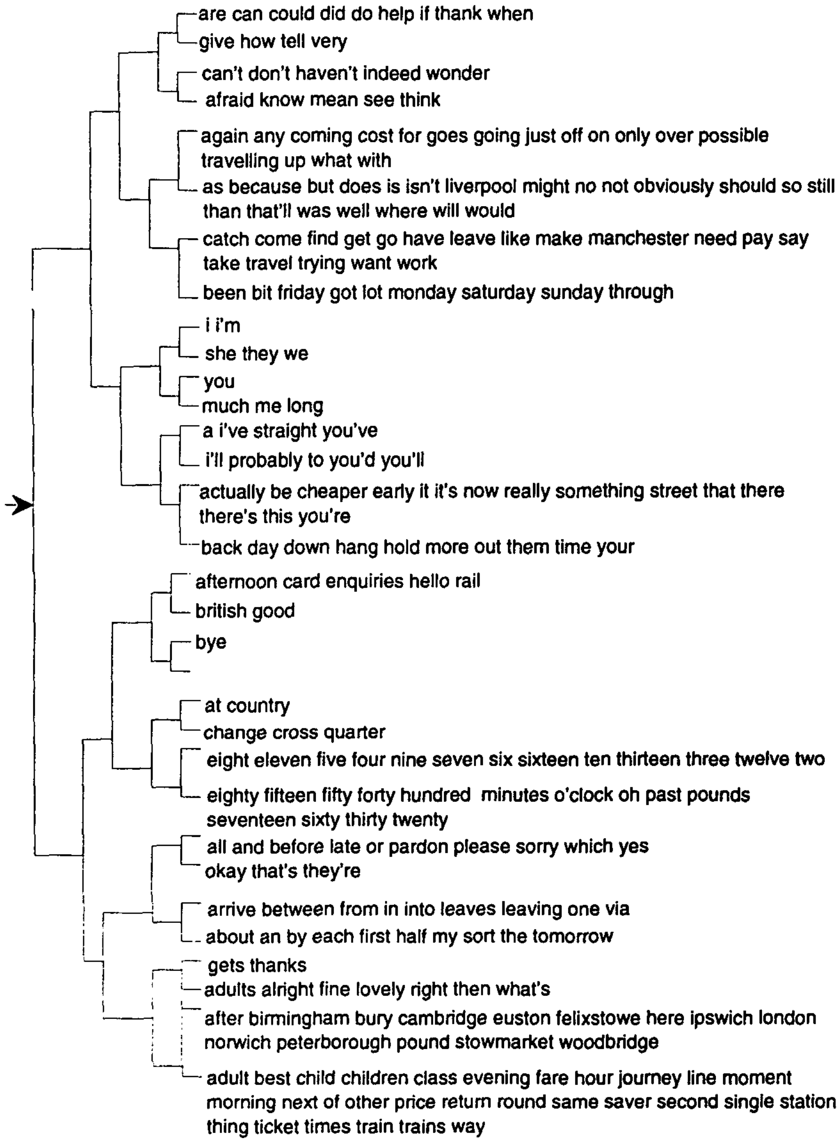

are can could did do help if thank when
give how tell very
can't don't haven't indeed wonder
afraid know mean see think
again any coming cost for goes going just off on only over possible travelling up what with
as because but does is isn't liverpool might no not obviously should so still than that'll was well where will would
catch come find get go have leave like make manchester need pay say take travel trying want work
been bit friday got lot monday saturday sunday through
i i'm
she they we
you
much me long
a i've straight you've
i'll probably to you'd you'll
actually be cheaper early it it's now really something street that there there's this you're
back day down hang hold more out them time your
afternoon card enquiries hello rail
british good
bye
at country
change cross quarter
eight eleven five four nine seven six sixteen ten thirteen three twelve two
eighty fifteen fifty forty hundred minutes o'clock oh past pounds seventeen sixty thirty twenty
all and before late or pardon please sorry which yes
okay that's they're
arrive between from in into leaves leaving one via
about an by each first half my sort the tomorrow
gets thanks
adults alright fine lovely right then what's
after birmingham bury cambridge euston felixstowe here ipswich london norwich peterborough pound stowmarket woodbridge
adult best child children class evening fare hour journey line moment morning next of other price return round same saver second single station thing ticket times train trains way

Figure 11
Top-down clustering. Classification of the most frequent words of a formatted VODIS corpus, using a top-down method.

rest of the vocabulary, including the verbs, to be distributed through the other half. Considering the distributions of pronouns and determiners, the merge-based system performs slightly better.

In conclusion, the two systems display the same kinds of differences and similarities as were seen when we compared our system to Elman's neural network—that is, our method performs slightly better with respect to overall classification topology, but loses in quality at lower levels. This loss in performance is also noted by Magerman (1994), who applies a binary classification tree to the task of parsing. Magerman also makes the point that trees (as opposed to directed graphs) are inherently vulnerable

Table 1
Reduced tag set used in Hughes-Atwell evaluation system.

ADJ	ADV	ART	CCON
CARD	DET	EX	EXPL
LET	MD	NEG	NOUN
ORD	OTH	PAST	PREP
PRES	PRON	PUNC	QUAL
SCON	TO	WH	

to unnecessary data fragmentation. The inaccuracies introduced by the first of these characteristics may be controlled, to a limited extent only, by using a hybrid top-down and bottom-up approach: instead of clustering vocabulary items from the top down, we could first merge some words into small word classes. Later top-down clustering would operate on these word groups as if they were words.

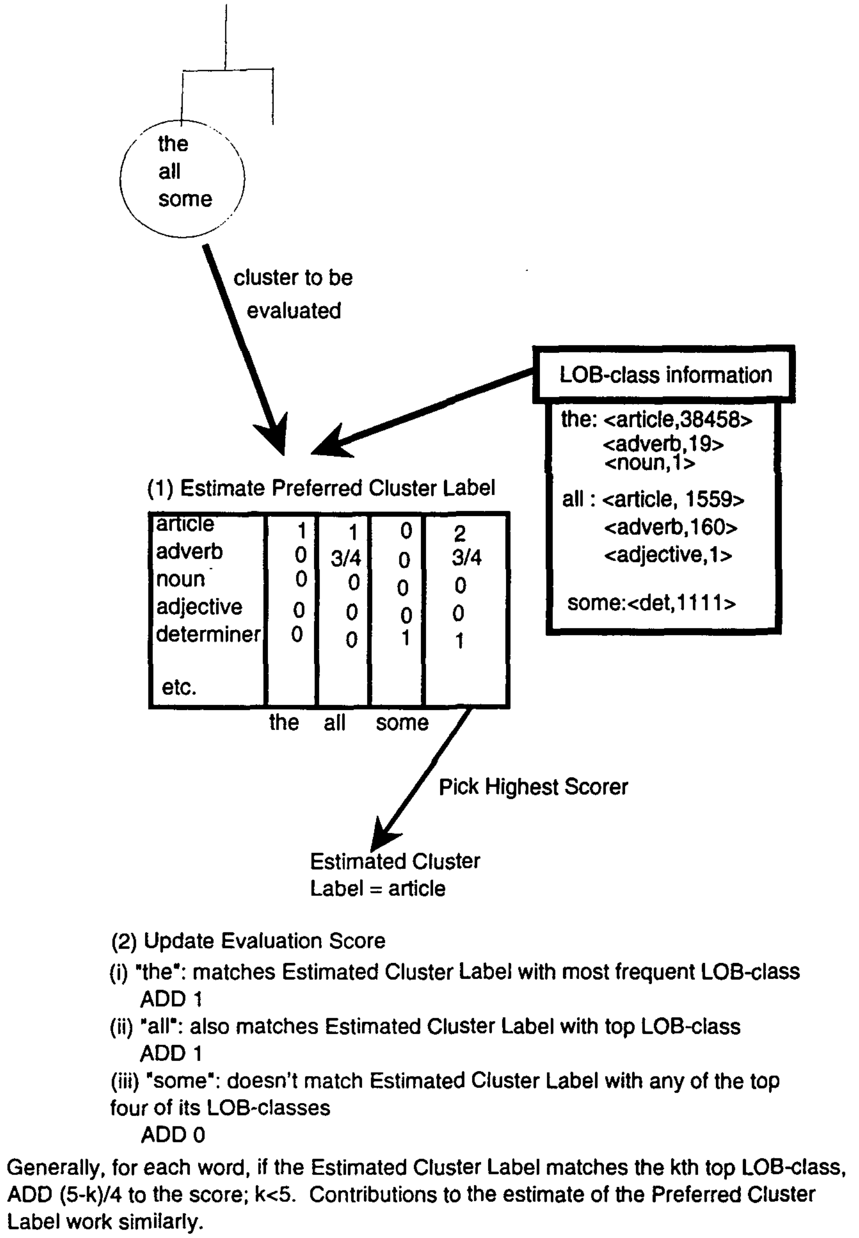
3.2.2 Quantitative comparison. Arriving at more quantitatively significant conclusions is difficult; Hughes (1994), for example, suggests benchmark evaluation—a standard tagged corpus (e.g., the LOB) is used as a reference against which automatic comparisons can be made. While this may not be appropriate for the designers of every automatic classification system, such as researchers whose main interest is in automatic classification in statistical language modeling, it has many advantages over qualitative inspection by an expert as an evaluation method, which to date has been the dominant method. Brill and Marcus (1992) suggest a similar idea for evaluating an automatic part-of-speech tagger.

Classification trees can be sectioned into distinct clusters at different points in the hierarchy; each of these clusters can then be examined by reference to the distribution of LOB classes associated to each word member of the cluster. A high-scoring cluster is one whose members are classified similarly in the tagged LOB corpus. In the following, we follow Hughes' method.

The evaluation is performed on the 195 most frequent words of the LOB corpus. The words are automatically classified using our top-down algorithm. The resulting classification is then passed to the evaluator, which works as follows: The first stage involves producing successive sections, cutting the tree into distinct clusters (from one cluster to as many clusters as there are vocabulary items), so that an evaluation score can be generated for each level; these evaluations can be plotted against the number of clusters. At each section, and for each cluster, we make an estimate of the preferred classification label for that cluster by finding the most common parts of speech associated with each word in the classification under question. For that part of speech most frequently associated with the word, we give a high weight, with decreasing weight for the second most frequent part of speech, and so on for the top four parts of speech. We then estimate the most likely part-of-speech category and label this cluster accordingly. Then, for each member of this cluster, a partial score is calculated that rates our classification of the word against its distribution of LOB classes. The summed score is then normalized as a percentage. An outline of the evaluation scheme is shown in Figure 12.

Hughes does not use the classification system provided with the LOB corpus—instead, he uses a reduced classification system consisting of 23 class tags, shown in Table 1. The results are shown in Figure 13.

Both of the compared classification systems use familiar statistical measures of correlation (Spearman's rank correlation coefficient and Manhattan metric) and grouping



	the	all	some	
article	1	1	0	2
adverb	0	3/4	0	3/4
noun	0	0	0	0
adjective	0	0	0	0
determiner	0	0	1	1
etc.				

(2) Update Evaluation Score

(i) "the": matches Estimated Cluster Label with most frequent LOB-class
ADD 1

(ii) "all": also matches Estimated Cluster Label with top LOB-class
ADD 1

(iii) "some": doesn't match Estimated Cluster Label with any of the top four of its LOB-classes
ADD 0

Generally, for each word, if the Estimated Cluster Label matches the kth top LOB-class, ADD (5-k)/4 to the score; k<5. Contributions to the estimate of the Preferred Cluster Label work similarly.

Figure 12
Hughes-Atwell Cluster Evaluation.

(group averaging and Ward's method) as their main method. Our system scores higher than the Finch system at all levels; the Hughes system scores better than ours over the first eighty classes, but worse at lower levels. However, we note that both Hughes and Finch use contiguous and noncontiguous bigram information, whereas we use contiguous bigram information only—the simplest estimate of word context possible—to explore just how much information could be extracted from this minimal context.

One limitation of Hughes' evaluation system is that fractured class distributions are not penalized: if some subbranch of the classification contains nothing but number

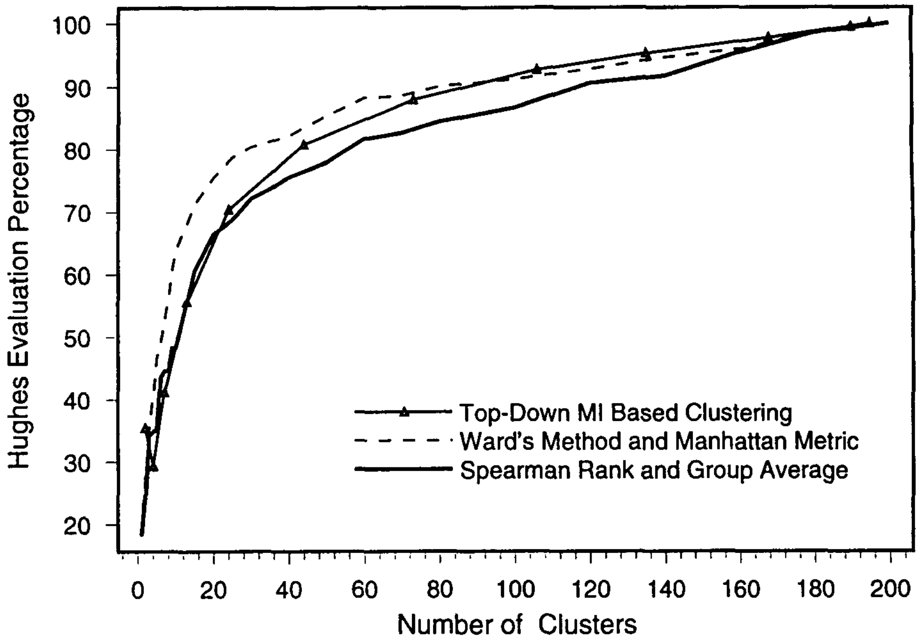


Figure 13
Performance comparison. Graph showing the performance of the top-down classification system compared to two recent systems — those of Hughes and Atwell and of Finch and Chater. Performance is measured by the Hughes-Atwell cluster evaluation system.

words (⟨ten⟩, ⟨three⟩, etc.) then that branch gets a certain score, regardless of how spread-out the words are within that branch. On the other hand, there may well be good engineering reasons to treat linguistically homogeneous words as belonging to different classes. For example, in a corpus of conversations about train timetables, where numbers occur in two main situations—as ticket prices and as times—we might expect to observe a difference between, say, the numbers from 1 to 12, and numbers up to 59 (hour numbers and minute numbers respectively); Figure 11 lends some support to this speculation. Similarly, phrases like ⟨ five pounds ninety nine pence ⟩ could lead to different patterns of collocation for number words. This sort of effect is indeed observed (McMahon 1994). It is less clear whether our main clustering result separates number words into different classes for the same kind of reason (in Figure 2, class 00000 contains 4 number words and class 00101 contains 11). A second limitation lies in the evaluation scheme estimating the canonical part of speech based on the rank of the parts of speech of each word in it—a better system would make the weight be some function of the probability of the parts of speech. A third criticism of the scheme is its arbitrariness in weighting and selecting canonical classes; the criticism is only slight, however, because the main advantage of any benchmark is that it provides a standard, regardless of the pragmatically influenced details of its construction.

Automatic word-classification systems are intrinsically interesting; an analysis of their structure and quality is itself an ongoing research topic. However, these systems can also have more immediate uses. The two types of use are related to the two types of approach to the subject—linguistic and engineering. Consequently, indirect evaluation can be linguistic or engineering-based.

Indirect linguistic evaluation examines the utility of the derived classes in solving various linguistic problems: pronoun reference (Elman 1990; Fisher and Riloff

1992), agreement, word-sense disambiguation (Liddy and Paik 1992; Gale, Church, and Yarowsky 1992; Yarowsky 1992; Pereira, Tishby, and Lee 1993) and resolution of anaphoric reference (Burger and Connolly 1992). A classification is said to be useful if it can contribute to a more accurate linguistic parse of given sentences. If our main interest were linguistic or cognitive scientific, we would be even more concerned about the way our system cannot handle multimodal word behavior and about the resulting misclassifications and fracturing of the classes.

One main engineering application that can use word classes is the statistical language model. Classifications which, when incorporated into the models, lower the test set perplexity are judged to be useful.

4. Structural Tags in Multiclass Statistical Language Models

There are several ways of incorporating word-classification information into statistical language models using the structural tag representation (McMahon 1994). Here, we shall describe a method, derived from Markov model theory (Jelinek and Mercer 1980), which is based on interpolating several language components. The interpolation parameters are estimated by using a held-out corpus. We decided to build an interpolated language model partly because it has been well studied and is familiar to the research community and partly because we can examine the lambda parameters directly to see if weight is indeed distributed across multiple class levels. A poor language model component will receive virtually no weight in an interpolated system—if we find that weight is distributed mostly with one or two components, we can conclude that interpolated language models do not find much use for multiple class information.

For the following experiments, a formatted version (punctuation removed, all words decapitalized, control characters removed) of the one-million-word Brown corpus was used as a source of language data; 60% of the corpus was used to generate maximum likelihood probability estimates, 30% to estimate frequency-dependent interpolation parameters, and the remaining 10% as a test set. The vocabulary items extracted from the training set were clustered according to the method described earlier.

For comparison, we calculated some test set perplexities of other language models. Improved performance can be obtained by making interpolation parameters depend upon some distinguishing feature of the prediction context. One easily calculated feature is the frequency of the previously processed word. In our main experiment, this resulted in 428 sets of λ values, corresponding to 428 different previous-word frequencies. The parameters are fitted into an interpolated language model the core of which is described by the equation:

$$P(w_k) = \lambda_u(f) \times P(w_k) + \lambda_b(f) \times P(w_k \mid w_{k-1}) + \lambda_t(f) \times P(w_k \mid w_{k-2}, w_{k-1})$$

where $f = f(w_j)$, the frequency of word $\langle w_j \rangle$ if a valid w_j exists and 0 otherwise—namely at the beginning of the test set, and when the previous word is not in the training vocabulary. The λ values are selected using a standard re-estimation algorithm (Baum et al. 1970). The resulting perplexity value for this system is 621.6. This represents a pragmatically sensible baseline value against which any variant language model should be compared. A similar word-based language model, the weighted average language model, has been developed by O'Boyle, Owens, and Smith (1994). This model is described as follows:

$$P(w_k \mid w_1^{k-1}) = \frac{\sum_{i=1}^{m} \lambda_i \times P_{ML}(w_k \mid w_{k-i}^{k-1}) + \lambda_0 \times P_{ML}(w_k)}{\sum_{i=0}^{m} \lambda_i}$$

where there are statistically significant segments up to $m+1$ words long and $P_{ML}(w_k)$ is the maximum likelihood probability estimate of a word. The numerator acts as a normalizer. It has been found that:

$$\lambda_i = 2^{(|w_{k-i}^{k-1}|)} \times \log f(w_{k-i}^{k-1})$$

where $|w_{k-i}^{k-1}|$ is the size of the segment, results in a useful language model of this form. When applied to the Brown corpus, excluding the 30% allocated for interpolation and only using n-grams up to 3, the model still performs well, achieving a perplexity score of 644.6; adding the extra training text should remove the disadvantage suffered by the weighted average model but at the probable cost of introducing new vocabulary items, making the test set perplexity comparisons even more difficult to interpret.

An important component of many statistical language-modeling systems is the bigram conditional probability estimator $\hat{P}(w_i \mid w_{i-1})$ (Church and Gale 1991); we shall restrict our attention to the case where both words have been seen before, though the bigram $\langle w_{i-1}, w_i \rangle$ itself may be unseen. We shall suggest an alternative to the familiar maximum likelihood bigram estimate, which estimates the probability as $\hat{P}(w_i \mid w_{i-1}) = \frac{f(w_{i-1},w_i)}{f(w_{i-1})}$, where $f(w)$ is just the frequency of occurrence of w in some training corpus.

The general form of the multilevel smoothed bigram model is:

$$\hat{P}(w_i \mid w_{i-1}) = \sum_{s=1}^{S} \lambda_{\phi(w_{i-1}),s} P(C^s(w_i) \mid C^s(w_{i-1})) P(w_i \mid C^s(w_i)) \quad (3)$$

where there are S levels of class granularity and $C^s(w_i)$ is the class at level s of the word w_i; $\lambda_{\phi(w_{i-1}),s}$ is an interpolation weight for language model component s and depends upon some function $\phi(w_{i-1})$ of the conditioning word w_{i-1}; common functions include a frequency-based interpolation $\phi(w_{i-1}) = f(w_{i-1})$ and a depth-s class-based interpolation, $\phi(w_{i-1}) = C^s(w_{i-1})$, though ϕ can partition the conditioning context in any way and this context does not necessarily have to be a recent word. The λ values are estimated as before, using the frequency of the previous word to partition the conditioning context. Parameter setting for the smoothed bigram takes less than a day on a Sparc-IPC.

Our 16-bit structural tag representation allows us to build an interpolated bigram model containing 16 levels of bigram-class information. As suggested earlier, we can look at the spread of λ values used by the smoothed bigram component as a function of the class granularity and frequency of the conditioning word. Figure 14 shows clearly that the smoothed bigram component does indeed find each class level useful, at different frequencies of the conditioning word. Next, we need to find out how much of an improvement we can achieve using this new bigram model.

We can replace the maximum likelihood bigram estimator in our interpolated trigram model with the smoothed bigram estimator. When we do, we get a perplexity of 577.4, a 7.1% improvement on standard interpolation, which scores 621.6. Other experiments with λ depending on the class (at a certain depth) of the previous word lead to smaller improvements and are not reported here.

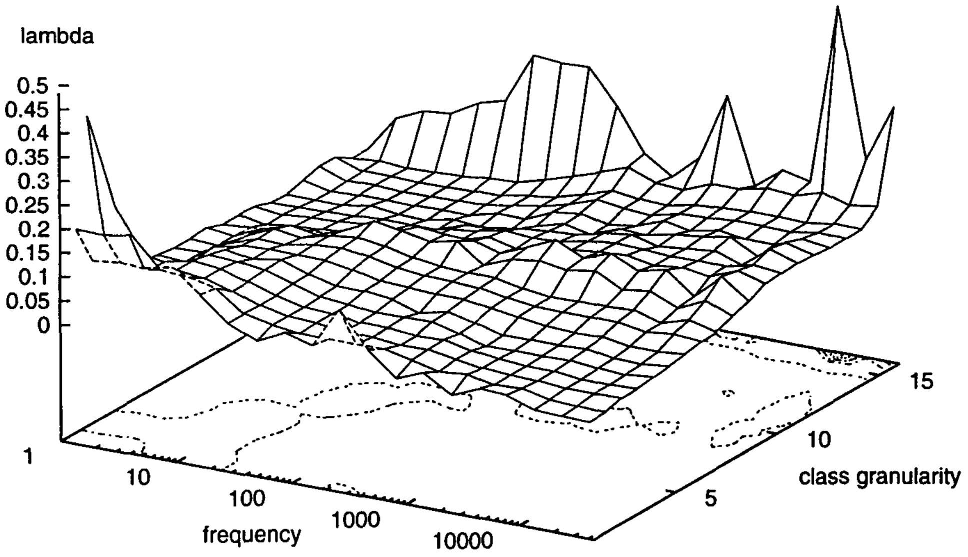


Figure 14
Bigram lambda weights. Surface showing how lambda varies with frequency (log scale) of previous word and bigram-class granularity. The projected contour map highlights the main feature of this relationship—at various frequencies, each of the 16 class bigram models is used.

Figure 15 summarizes the test set perplexity results. We note that our 7.1% improvement is larger than that obtained by Brown, Della Pietra, DeSouza, Lai, and Mercer (1992), who report a 3.3% improvement. The smaller absolute perplexity scores they quote are a consequence of the much larger training data they use. One reason for this apparent improvement may be that their baseline model, constructed as it is out of much more training data, is already better than our equivalent baseline, so that they find improvements harder to achieve. Another reason may be due to the different vocabulary sizes used (Ueberla 1994). A third reason, and one which we consider to be important, is that multilevel class-based language models may perform significantly better than two-level ones. We carried out another experiment to support this claim.

We constructed a frequency-dependent interpolated unigram and bigram model as a baseline. Its test set perplexity was 635. We then replaced the maximum likelihood bigram component with the smoothed bigram estimate. The perplexity for this system was 580, a 9% improvement. We also replaced the maximum likelihood bigram component with a series of 15 two-level smoothed bigram models—from a 16-plus-15 smoothed bigram to a 16-plus-1 smoothed bigram. Figure 16 details these results. The best of these two-level systems is the 16-plus-8 model, which scores 606. So, on a bigram model, the multilevel system is 4.3% better than the best two-level system, which supports our claim. We chose bigram models in this experiment so that we could make some comparisons with similarity-based bigram models.

Dagan, Markus, and Markovitch (1993) claim that word-classification systems of this type may lead to substantial information loss when compared to similarity methods (Dagan, Pereira, and Lee 1994; Essen and Steinbiss 1992). The similarity-based system of Dagan, Pereira, and Lee (1994) improves a baseline Turing-Good bigram model by 2.4% and the co-occurrence system of Essen and Steinbiss (1992) leads to a 10% improvement over an interpolated baseline bigram model. This latter result is based on a similarly sized training set and so our 9% improvement compared to their

Language Model	Test Set Perplexity
Weighted Average	644.626
Interpolated Trigram	621.632
Interpolated Trigram (smoothed bigram component)	577.421

Figure 15
Test set perplexity improvements. When an interpolated trigram language model uses smoothed bigram estimates, test set perplexity reduced by approximately 7.1% compared to a similar system with maximum likelihood bigram estimates, and 10% compared to the weighted average language model.

Language Model	Test Set Perplexity
Baseline bigram	635
word plus class 1	634
word plus class 2	633
word plus class 3	626
word plus class 4	621
word plus class 5	616
word plus class 6	614
word plus class 7	609
word plus class 8	606
word plus class 9	609
word plus class 10	614
word plus class 11	618
word plus class 12	622
word plus class 13	627
word plus class 14	631
word plus class 15	633
Multilevel	580

Figure 16
Multilevel versus two-level bigram performances. A multilevel smoothed bigram model is 9% better than a baseline maximum likelihood model and 4.3% better than the best two-level class-based bigram model.

10% suggests that language models based upon fixed-place classes can be only slightly worse than some similarity models, given approximately equal training texts.

4.1 An Example

As an illustration of the kind of advantage structural tag language models can offer, we introduce nine oronyms (word strings which, when uttered, can produce the same sound) based upon the uttered sentence:

The boys eat the sandwiches.

If we assume that we already possess a perfect speech recognition acoustic model (Jelinek, Mercer, and Roukos 1992), it may be able to recover the phoneme string:

/DH a b OI z EE t DH A s AA n d w i j i z/

Sentence	W.A.	Smoothed	Grammatical
the boy seat the sandwiches	3,419	7,848	no
the boys eat the sandwiches	1,787	8,821	yes
the boy seat this and which is	435	137	no
the boys eat this and which is	232	149	no
the buoys eat the sandwiches	195	469	yes
the buoys eat this and which is	25	8	no
the boys eat the sand which is	14	21	yes
the buoys eat the sand which is	1.5	1.1	yes
the buoy seat this and which is	0	0	no

Figure 17
Improvements in a simulated speech-recognition example. Nine versions of a phonemically identical oronym, ordered by weighted average (W.A.) probability ($\times 10^{-20}$). The W.A. language model ranks the preferred sentence second. The smoothed structural tag model successfully predicts the original utterance as the most likely. ⟨buoy⟩ is an unseen vocabulary item in this test. Also, in all but two nonzero cases, the smoothed model makes grammatically correct sentences more likely and vice versa.

The original sentence is not the only speech utterance that could give rise to the observed phoneme string; for example, the meaningless and ungrammatical sentence:

*The buoy seat this and which is.

can also give rise to the observed phonemic stream. Humans usually reconstruct the most likely sentence successfully, but artificial speech recognizers with no language model component cannot. Nonprobabilistic models, while theoretically well-grounded, so far tend to have poor coverage. Another limitation can be seen if we consider a third hypothesized sentence:

The buoys eat the sand which is.

This simultaneously surreal and metaphysical sentence may be accepted by grammar systems that detect well-formedness, but it is subsequently considered just as plausible as the original sentence. A probabilistic language model should assign a relatively low probability to the third sentence. We constructed nine hypothesized sentences, each of which could have produced the phoneme string; we presented these sentences as input to a high-quality word-based language model (the weighted average language model) and to another smoothed structural tag language model. Neither the Hughes system nor the Finch system are ever applied to language models; also, the details of the Brown language model are insufficient for us to rebuild it and run our sentences through it. Figure 17 shows the normalized probability results of these experiments. The new language model successfully identifies the most likely utterance. In all but two nonzero cases, grammatically well-formed sentences are assigned a higher raw probability by the new model, and vice-versa for ungrammatical sentences.

Using the top two sentences ⟨the boy seat the sandwiches⟩ and ⟨the boys eat the sandwiches⟩, we can examine the practical benefits of class information for statistical language modeling. An important difference between the two is in the bigrams ⟨boy seat⟩ and ⟨boys eat⟩, neither of which occurred in the training corpus. The model that uses word frequencies exclusively differentiates between the two hypothesized sentences by examining the unigrams ⟨boy⟩, ⟨seat⟩, ⟨boys⟩, and ⟨eat⟩. In our training corpus, ⟨boy⟩ and ⟨seat⟩ are individually more likely than ⟨boys⟩ and ⟨eat⟩. However, with the structural tag model, extra word-class information allows the system to prefer the more common noun-verb pattern. This sort of advantage becomes even more apparent with number words: for example, if we were trying to predict the likelihood of ⟨seconds⟩ given ⟨six⟩, even though the bigram ⟨six seconds⟩ does not occur in our training text, we find that ⟨three seconds⟩, ⟨four seconds⟩, and ⟨five seconds⟩ occur, as do ⟨six years⟩, ⟨six months⟩, ⟨six weeks⟩, and ⟨six days⟩.

5. Discussion

The automatic word-classification system based on a binary top-down mutual information algorithm leads to qualitatively interesting syntactic and semantic clustering results; quantitatively, it fares well compared with other systems, demonstrating complementary strengths and weaknesses compared to the more usual merge-based classification systems. Results from an implementation of one version of a multilevel class-based language model (an interpolated trigram model with the maximum likelihood bigram component replaced with a smoothed bigram component) show a 7% improvement in statistical language model performance compared to a standard interpolated language model. We have incorporated structural tag information into an interpolated model because it provides a well-attested and successful base system against which improvement can be measured; it also offers us the opportunity to visualize the λ distribution across 16 classes so that we can observe in which circumstances each class level is preferred (see Figure 14). However, we believe that the weighted average system described earlier, with its scope for improvements including n-gram information beyond the trigram and its avoidance of data-intensive and computationally intensive parameter optimization, could offer a more convenient platform within which to place structural tag information. Although variable granularity class-based language models will never fully capture linguistic dependencies, they can offer modest advances in coverage compared to exclusively word-based systems.

Acknowledgments
Both authors thank the Oxford Text Archive and British Telecom for use of their corpora. The first author wishes to thank British Telecom and the Department of Education for Northern Ireland for their support.

References

Bahl, Lalit R., Peter F. Brown, Peter V. DeSouza, and Robert L. Mercer. 1989. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(7):1001–1008, July.

Bahl, Lalit R., Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analyses of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Beckwith, Richard, Christiane Fellbaum, Derek Gross, and George A. Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, chapter 9, pages 211–232.

Black, Ezra, Roger Garside, and Geoffrey Leech. 1993. *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach*. Rodopi.

Black, Ezra, Frederick Jelinek, John Lafferty, David M. Magerman, Robert Mercer, and Salim Roukos. 1993. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 31–37, Ohio State University, June.

Brill, Eric, David Magerman, Mitchell Marcus, and Beatrice Santorini. 1990. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop.*

Brill, Eric, and Mitch Marcus. 1992. Tagging an unfamiliar text with minimal human supervision. In *Probabilistic Approaches to Natural Language.* American Association for Artificial Intelligence, AAAI Press. Technical report FS-92-05.

Brown, Peter F., Vincent Della Pietra, Peter DeSouza, Jennifer C. Lai, and Robert Mercer. 1992. Class-based n-gram models of natural langauge. *Computational Linguistics,* 18(4):467–479.

Brown, Peter F., Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics,* 18(1):31–40.

Burger, John D. and Dennis Connolly. 1992. Probabilistic resolution of anaphoric reference. In *Probabilistic Approaches to Natural Language.* American Association for Artificial Intelligence, AAAI Press. Technical report FS-92-05.

Church, Kenneth Ward. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing.*

Church, Kenneth W. and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language,* 5:19–54.

Church, Kenneth W., William A. Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates, chapter 6, pages 115–164.

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics,* 19(1):1–23.

Cookson, S. 1988. Final evaluation of VODIS. In *Proceedings of Speech '88, Seventh* FASE *Symposium,* pages 1311–1320, Edinburgh. Institute of Acoustics.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory.* John Wiley and Sons.

Dagan, Ido, Shaul Markus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the Association for Computational Linguistics,* pages 164–171.

Dagan, Ido, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurence probabilities. In *Proceedings of the Association for Computational Linguistics.*

Derouault, Anne-Marie and Bernard Merialdo. 1986. Natural language modelling for phoneme-to-text transcription. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence,* PAMI-8(6), November.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science,* 14:179–211.

Essen, Ute and Volker Steinbiss. 1992. Co-occurrence smoothing for stochastic language modelling. In *Proceedings of ICASSP,* volume 1, pages 161–164.

Finch, Steven and Nich Chater. 1991. A hybrid approach to the automatic learning of linguistic categories. *A.I.S.B. Quarterly.*

Finch, Steven and Nick Chater. 1992. Bootstrapping syntactic categories using statistical methods. In Walter Daelemans and David Powers, editors, *Background and Experiments in Machine Learning of Natural Language,* pages 229–235. Institute for Language Technology and AI.

Finch, Steven and Nick Chater. 1994. Learning syntactic categories: A statistical approach. In M. Oaksford and G.D.A. Brown, editors, *Neurodynamics and Psychology.* Academic Press, chapter 12.

Fisher, David and Ellen Riloff. 1992. Applying statistical methods to small corpora: Benefitting from a limited domain. In *Probabilistic Approaches to Natural Language.* American Association for Artificial Intelligence, AAAI Press. Technical report FS-92-05.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *probabilistic Approaches to Natural Language.* American Association for Artificial Intelligence, AAAI Press. Technical report FS-92-05.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika,* 40:237–264, December.

Hughes, John. 1994. *Automatically Acquiring a Classification of Words.* Ph.D. thesis, School of Computer Studies, University of Leeds.

Hughes, John and Eric Atwell. 1994. The automated evaluation of inferred word classifications. In *Eleventh European Conference on Artificial Intelligence.*

Jelinek, Frederick. 1976. Continuous speech recognition by statistical methods.

Proceedings of the I.E.E.E., 64(4), April.

Jelinek, Frederick. 1985. The development of an experimental discrete dictation recogniser. *Proceedings of the I.E.E.E.*, 73(11).

Jelinek, Frederick and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam.

Jelinek, Frederick, Robert L. Mercer, and Salim Roukos. 1990. Classifying words for improved statistical language models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 621–624, Albuquerque, New Mexico.

Jelinek, Frederick, Robert L. Mercer, and Salim Roukos. 1992. Principles of lexical language modelling for speech recognition. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*. Maral Dekku, Inc.

Johansson, Stig J., Eric S. Atwell, Roger Garside, and Geoffrey Leech. 1986. *The Tagged LOB Corpus: User's Manual*. The Norwegian Centre for the Humanities, Bergen.

Katz, Slava M. 1987. Estimation of probabilities for sparse data for the language model component of a speech recogniser. *I.E.E.E. Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401, March.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680, May.

Kiss, George R. 1973. Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, 7:1–41.

Kneser, Reinhard and Hermann Ney. 1993. Forming word classes by statistical clustering for statistical language modelling. In R. Köhler and B. B. Rieger, editors, *Contributions to Quantitative Linguistics*. Kluwer Academic Publishers, pages 221–226.

Kuhn, Ronald and Renato De Mori. 1990. A cache-based natural language model for speech recognition. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, June.

Kupiec, Julian. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6:225–242.

Liddy, Elizabeth D. and Woojin Paik. 1992. Statistically-guided word sense disambiguation. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press. Technical report FS-92-05.

Magerman, David M. 1994. *Natural Language Parsing as Statistical Pattern Recognition*. Ph.D. thesis, Stanford University Computer Science Department, February.

Matsukawa, Tomoyoshi. 1993. Hypothesizing word association from untagged text. In *ARPA Workshop on Human Language Technology*, Princeton, March.

McMahon, John. 1994. *Statistical Language Processing Based on Self-Organising Word Classification*. Ph.D. thesis, Department of Computer Science, Queen's University of Belfast.

McMahon, John and F. J. Smith. 1994. Structural tags, annealing and automatic word classification. *Artificial Intelligence and the Simulation of Behaviour Quarterly*, 90.

Ney, Hermann, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38.

O'Boyle, Peter, Marie Owens, and F. J. Smith. 1994. A weighted average N-gram model of natural language. *Computer Speech and Language*, 8:337–349.

Pereira, Fernando and Naftali Tishby. 1992. Distributed similarity, phase transitions and hierarchical clustering. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press. Technical report FS-92-05.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the Association for Computational Linguistics*, pages 183–190.

Rabiner, L. R. and B. J. Juang. 1986. An introduction to hidden Markov models. *I.E.E.E. A.S.S.P. Magazine*, pages 4–16, January.

Redington, Martin, Nick Chater, and Steven Finch. 1993. Distributional information and the acquisition of linguistic categories: A statistical approach. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*.

Redington, Martin, Nich Chater, and Steven Finch. 1994. The potential contribution of distributional information to early syntactic category acquisition. Unpublished Report.

Resnik, Philip S. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, Computer and Information Science, University of Pennsylvania, December.

Institute for Research in Cognitive Science Report I.R.C.S.-93-42.

Sampson, Geoffrey. 1987. Evidence against the grammatical/ungrammatical distinction. In Wilem Meijs, editor, *Corpus Linguistics and Beyond—Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora.* Rodopi, Amsterdam, pages 219–226.

Schütze, Hinrich. 1993. Part-of-speech induction from scratch. In *Proceedings of the Association for Computational Linguistics 31*, pages 251–258.

Schütze, Hinrich. 1995. Distributional part-of-speech tagging. In *Proceedings of the Seventh European Chapter of the Association for Computational Linguistics,* University College Dublin, March.

Ueberla, Joerg. 1994. Analysing a simple language model—some general conclusions for lanaguage models for speech recognition. *Computer Speech and Language,* 8:153–176.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, pages 454–460.

Zipf, George K. 1949. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley.