A Review of Statistical Language Processing Techniques

J. McMahon[*] and *F.J. Smith*
The Queen's University of Belfast

February 8, 1995

Abstract

We present a review of some recently developed techniques in the field of natural language processing. This area has witnessed a confluence of approaches which are inspired by theories from linguistics and those which are inspired by theories from information theory: statistical language models are becoming more linguistically sophisticated and the models of language used by linguists are incorporating stochastic techniques to help resolve ambiguities. We include a discussion about the underlying similarities between some of these systems and mention two approaches to the evaluation of statistical language processing systems.

1 Introduction

Within the last decade, a great deal of attention has been paid to techniques for processing large natural language copora. The purpose of much of this activity has been to refine computational models of language so that the performance of various technical applications can be improved (*e.g.* speech recognisers [67], speech synthesisers [32], optical character recognisers [65], lexicographical support tools [29], automatic translation systems [21] and information retrieval and document analysis systems [29]); another significant interest is shown by cognitive scientists who build explicit computational models of the human language processing ability [76, 26]. These two sets of interests are not necessarily mutually exclusive.

Already, several sub-domains have crystallised out of the current surge of interest: automatic word classification, automatic part-of-speech tagging, segmentation of streams of linguistic units (at the sentence, phoneme, morpheme and word level), language models for recognition (in this substantial research domain, mostly untouched by cognitive scientific concerns, the phrase 'language model' has become synonymous with the linguistically discredited 'finitary model' which, with the addition of stochastic transition arcs, is also known as a 'Markov model'), grammar induction and machine translation. We shall examine some of the work which has been carried out in these areas and also situate the efforts within the broader context of models of natural language.

We shall also be looking at the inventory of tools and techniques which researchers have been applying to the various areas of language processing: these include entropy, perplexity, mutual information, traditional statistics, connectionism, genetic algorithms, formal language theory, Markov modelling and non-linear dynamical modelling. After this, we suggest that there are interesting mathematical connections between many of these techniques. We finish by discussing two ways of evaluating language modelling systems: engineering evaluations and cognitive scientific evaluations.

[*]Department of Computer Science, Q.U.B., Belfast BT7 1NN, N. Ireland. Email: J.McMahon@qub.ac.uk

2 Corpus Processing Specialisms

In this section we describe some examples of recent work within each of the new language processing specialisms. We present this selection as being representative of the main approaches, rather than offering an exhaustive catalogue. Whilst we believe that this approach is more than merely expository, it is clear that there are many areas of overlap both in terms of the work which particular researchers undertake and in terms of mathematical underpinnings of the specialisms. Section 3 explores these connections in more detail and we shall now offer some general remarks on computational models of language.

Two traditionally important concepts in linguistics are syntagmatic and paradigmatic relations [35]. The former define ways of combining linguistic units and the latter define similarities between linguistic units, though these relations are interdependent. Of our specialisms, we associate word classification with the paradigmatic relation; segmentation, on the other hand, is more immediately associated with the syntagmatic relation. However, we shall see that word classification makes little sense without syntagmatic considerations; a similar inter-connection holds in the process of segmentation. Within the field of language modelling for recognition, the syntagmatic relation has dominated early research, which was based upon n-gram distributions of words. More recently, class-based n-gram language models have improved language model performance precisely by incorporating paradigmatic information [20, 92]. The specialism of grammar induction, however, must derive both relations simultaneously — this task is sometimes called the bootstrapping problem [109]. The theme of the present work is to indicate the myriad ways that syntagum and paradigm can reduce uncertainty.

The interdependence of syntagmatic and paradigmatic relations in the structure of natural language is accepted by many corpus processing researchers [126, 47] — it is this interdependence which leads critics to discount the possibility of automatic natural language learning: in order to construct a system which generates a hierarchical structure from plain sentences, for example, one first needs to know to which categories individual words belong; but to know this, one must have some idea of the positions these words occupy in a grammar. Finch and Chater [47] situate this linguistic problem in the wider cognitive context of learning new domains.

Many of the researchers discussed in this review support the modern structural linguistic approach, which suggests that a significant amount of the structure of natural language can be detected by distributional and statistical means; Tanenhaus [130] summarises the search by early structural linguists for discovery procedures, which when applied mechanically to a corpus of utterances could, in principle, extract the linguistically relevant units. Liberman [84] notes the growing interest of the academic community in automatic approaches; this can be read as an indirect barometer of the success of these systems. Church and Mercer [31] argue that many of the techniques described herein are out-performing more traditional knowledge-based methods. The case for automatic modelling of linguistic phenomena over manual modelling is made convincingly in Makhoul *et al.* [88], which also discusses some of the practical requirements and problems associated with automatic linguistic modelling. Zernik [135] discusses many of the major limitations of current automatic language processing systems. He presents a cogent argument in favour of systems which need minimal human intervention and which process raw materials which are easy to construct.

Sampson [119] and Brill *et al.* [18] both present a strong case in favour of distributional analysis. Church [32] also finds surface statistics worth investigating; others chose to make no direct challenge to linguistic orthodoxy while using methods which undermine some key tenets of theoretical linguistics (Carroll and Charniak [24], for example consider phrases 'good' if they occur frequently, regardless of what linguists think; also, Magerman [87] declares in the preface of his thesis dissertation that a

significant goal of his work was to replace linguistics with statistical analysis of corpora).

Investigating the amount of linguistic structure in language utterances is an interesting theoretical research topic in itself, though it also commands practical advantages over a reliance on manually constructed corpora. First and most obviously, running an algorithm on raw text (for example, to generate word classes) is time and resource efficient — manually tagged corpora are expensive to make, largely because the humans who make the word class judgements do so slowly. Secondly, manual tagging is not a language-independent process whereas the same automatic word classification system could be applied to any language — even one whose syntax and possibly semantics are unknown to investigators. Finally, in some cases, automatic word classification may be the only method available to the researcher. Similar advantages apply to automatic part of speech tagging, automatic segmenting and parsing, and grammar induction.

Church and Hanks [33] have recently introduced the psycholinguistic term *word association* into the vocabulary of computational linguistics. In psycholinguistics the term refers to the usually semantic priming which occurs between pairs of words. Tests can be performed which measure the lexical retrieval speed for a word like ⟨`doctor`⟩; these tests can be repeated when the subject has been primed by being shown the word ⟨`nurse`⟩, for example. Those primer words which lower retrieval time are said to have a high word association index. Also, syntactically close words can act as primers — for example, between certain verbs and prepositions. These association indices are estimated through psycholinguistic experiments with many subjects (for example, see Miller and Charles [96])

The insight that semantic and syntactic relations could be induced from the linear structure of natural language utterances has long been a key tenet of structural linguistics [60, 130], crystallised by Firth [49] as follows: "You shall know a word by the company it keeps." Using information theory, the Firthian slogan can be re-cast: "The structural description of a lexical item is some function of its context". In other words, useful models of word context can lead to low entropy word prediction systems.

Only recently have computational resources of sufficient power and corpora of sufficient size been made available to the research community [31, 84] to allow them to perform some of the many experiments indicated by a structuralist perspective.

Word associations can be extracted from corpora by borrowing the information theoretic measure of *mutual information* [67, 69, 43, 34]; if $P(x)$ and $P(y)$ are the independent probabilities of events x and y, then the mutual information, $M(x, y)$ is

$$M(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \tag{1}$$

This measure compares how likely x and y are to occur together — in the case of words, this means serial occurrence, so that $M(x, y)$ is not necessarily the same as $M(y, x)$ — with their independent likelihoods of occurrence. The higher the likelihood of the co-occurrence of events x and y, the larger the mutual information value. Church and Hanks describe some initial analyses of corpora using mutual information. Most of their results are close examinations of particular word relations and syntactic constructions.

Bod [13] favours models which deal with language performance over competence. He describes four limitations of the competence approach: the problem of ambiguity proliferation, the instability of human grammaticality judgements, the poor facility for modelling language change and the general descriptive inadequacy of all existing rule-based grammars. This last problem has a tendency to become more limiting the larger a linguistically designed grammar gets — the more rules and features, the more chances for inappropriate interactions between them. This insight is analogous to one made

by many critics of traditional Artificial Intelligence methodologies [61, 10, 19, 131]. Gorin *et al.* [58] model the relationship between linguistic behaviour and situated meaning by mapping language input machine output, in a restricted domain. This approach shares some similarities with transfer-based approaches to machine translation (see 2.7). The competence grammar of a language user relates to the general structural capacities of that grammar and language, but by itself, it tells us nothing much about the details of how communities of language users have certain linguistic expectations and preferences and how these are used, practically, in disambiguating possibly confusing messages.

Bod predicts that the most useful language processing systems will be hybrids of the statistical and formal approaches. Resnik [117], Derouault *et al.* [37] and Solomon [125] also notice a coming together of information-theoretic and traditional linguistic approaches to language. Resnik's system uses a database of words which have been tagged using a semantic network structure. This database is input to a taxonomy-generating system whose principles are based on information theory. He suggests that this synthetic approach models language generation and understanding better than traditional linguistic approaches. He puts less emphasis on the debate about language acquisition and avoids using raw corpora. He re-states the familiar claims that lexically based statistics can only generate limited models of language competence.

2.1 Word Classification

We start our inventory of corpus based specialisms with word classification. We note first that in psycholinguistics, both of the major formalist approaches to language (Chomskyan [28] and Distributional [60]) need to postulate a system of lexical acquisition and categorisation; in the Chomskyan model, innate (universal) language structures alone do not provide information about nouns or verbs or, indeed, about the lexicon; distributionalists too need to explain by which mechanisms children discover word taxonomies. Corpus linguists have been aware of this semi-autonomy of lexical acquisition and language acquisition for at least two decades [76]. This situation allows researchers working on automatic word classification to remain neutral about full language acquisition. Of course, corpus processing researchers may be uninterested in cognitive scientific implications of their models: for example, if a linguistically naïve model of some aspect of human language processing nevertheless significantly improved some natural language application, some corpus processing researchers may consider the model successful; here, success is equated with utility.

Explicit information theoretic approaches to automatic word classification are common: Brown *et al.* [20], McMahon *et al.* [93], Pereira *et al.* [105] and Ney *et al.* [99] have used various measures taken from information theory as the bases of their systems. Connectionist networks are also well attested: Kiss [76], Elman [40, 41] and Finch *et al.* [47] present interesting systems, couched mainly in a cognitive scientific perspective. Finch *et al.* [46, 48] also investigate a word classification system based on a more traditional statistic — Spearman's rank sum correlation coefficient. Several other researchers have also resorted to standard statistical measures; these include Schütze [122, 121, 123] and Hughes *et al.* [63, 64]. Brill *et al.* [17] designed a word classification measure based explicitly on early structural linguistic theory.

The systems described in Brown *et al.* and McMahon *et al.* both make use of the mutual information between contiguous word-class pairs and work by a process of local optimisation. The mutual information between two random variables, X and Y is just a summed extention of equation 1

$$M(X,Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{2}$$

where $P(x)$ is the probability of of event x. When the events are occurrences of word classes, equation 2 measures how much more likely it is to observe the two classes in text than their independent unigram probabilities suggest. A score of 0 indicates that the observed class event $\langle \mathtt{xy} \rangle$ is no more common than the independent probabilities $P(x)$ and $P(y)$ suggest. A large positive score indicates that the event $\langle \mathtt{xy} \rangle$ is much more likely than the independent unigram probabilities suggest; and conversely for negative mutual information values. Brown embeds this metric in a bottom-up agglomerative algorithm the results of which can be used to produce a binary classification of the vocabulary. McMahon embeds the metric in a top-down algorithm which operates on *structural tag* representations of the vocabulary. Structural tags provide immediate and multi-level access to the classifications of every word in a vocabulary. Words are represented as n-bit numbers the most significant bits of which correspond to classifications of various granularities. Access to the n-bit word facilitates immediate access to its many classifications. Other classification systems implement a more explicitly functional relationship between a set of word-objects and a single set of class-objects. The clustering systems can be incorporated into an interpolated language model and successfully lowers the test set perplexity of such models.

The task of finding that series of class merges which maximises the class average mutual information has not been solved, though the technique of locally maximising average class mutual information does lead to very interesting results. Brown's algorithm discovered, among others, the classes:

```
'mother wife father son husband brother
 daughter sister boss uncle'
'had hadn't hath would've could've should've must've might've'
'head body hands eyes voice arm seat eye hair mouth'
```

McMahon *et al.* present some interesting semantic clustering results which were produced by analysing a hybrid tag-word version of the LOB corpus. Some classes include:

```
'arm breath breathing cheek chin coat eye fist forehead hair handkerchief ...'
'aunt brother father father-in-law husband mother sister uncle wife ...'
'decade year month fortnight week hour inch lot spot step ...'
'actor actress 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 soilder sovereign stranger teacher widow woman writer ...'
```

Pereira *et al.* use the Kullback-Leibler distance (relative entropy), $D(P||Q)$ to measure the distributional dissimilarities between word distributions P and hypothesised cluster centriods Q.

$$D(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{3}$$

The measure offers an indication of how much information loss would be incurred by using the distribution Q — based on a summarising distribution of many words — instead of the correct distribution P. They embed their measure in an algorithm based on temperature annealing: as the temperature constant is lowered, dissimilarities between the distributions become more important and the optimal number of cluster centriods increases. Non-hierarchical groups can be extracted from this process and intuitively appealing syntactic clusters are observed, including, for example, the group:

`quickly, apart, slowly, rapidly, quietly, shortly, sharply, steadily, remote, exclusively, softly, sadly, varies, eagerly .....`

Kneser *et al.* [77] use a maximum likelihood criterion to drive their word clustering, which they embed in an iterative optimisation algorithm. They postulate a mapping between words w and one of M word clusters: $w \rightarrow g(w)$, where $g(w)$ is the class of word w. They then use a training set to classify words in order to make the bigram class language model

$$P(w_1^n) = \prod_{i=1}^{n} P(g(w_i)|g(w_{i-1}))P(w_i|g(w_i)) \tag{4}$$

maximally likely (the sum is over all n words in the training set). In practice, they minimise the negative logarithm of this probability — called the estimated log-probability, LP. Some algebraic manipulation shows that LP is equivalent to

$$LP = -\sum_{g_1,g_2} N(g_1,g_2)\log N(g_1,g_2) + 2\sum_{g} N(g)\log N(g) - \sum_{w} N(w)\log N(w) \tag{5}$$

where $N(x)$ is the count of event x in the training set. The algorithm operates on an initial M-cluster where each of the $M-1$ most frequent words are in unique clusters and the remaining cluster contains all other words.

Elman introduces the recurrent neural network, in which input to the hidden layer of neurons is augmented with a copy of the states of the hidden neurons at the previous time interval. This allows some serial structure to be encoded in the network. After training the net to predict the next word in an unbroken sequence of sentences, Elman performs a cluster analysis of the hidden unit activation levels, from which he can generate a word taxonomy. This system discovers some interesting syntactic and semantic structure in an artificially generated simple grammar. His important insight was to recognise that after a network has been trained to predict words in a continuous stream from sentences generated by a simple grammar, the network's hidden nodes must be representing, in a distributed way, the syntactic-semantic distinctions of that language. Two aspects of this system make it difficult to use directly in statistical language modelling — first, the classes are not explicitly available and second, the quality of results when scaled up to real language data has been challenged [112].

Schütze takes the 5,000 most frequent words from a large corpus and generates a 5000 by 4 sparse matrix containing the relevant bigram frequencies. This matrix is passed to a sparse matrix algorithm which implements a singular value decomposition. This produces a 15-dimensional real-valued matrix for each word preserving similarities between words.

An interesting statistical model is that of Finch and Chater [46, 47]. They use a simple statistical measure to derive syntactic and some semantic categories. They derive their similarity metric from a consideration of the 'replacement test' of theoretical linguistics, which suggests that lexical items which are distributed similarly should receive similar linguistic categorisations. If $\langle C, w\rangle$ is a well-formed sentence in a language, where C is the set of all contexts and w is a particular lexical item, then w and w' are said to belong to the same class if $\langle C, w'\rangle$ is also well-formed. Since the notion of well-formedness is not simply incorporated into statistical natural language processing systems, Finch and Chater define the context of an item to be the two words either side of the word. They use the Spearman Rank Correlation Coefficient and cluster analysis then places words of similar distribution close to each other in a dendrogram.

One of their experiments involves the four positions (two either side of the word in question) as four vectors, each of 150 dimensions, corresponding to the frequency of the 150 most common words in

the corpus. These four vectors define a simplified and computationally tractable operational definition of context. For every word, w_i, a cluster of 600-dimensional points is calculated; the Spearman's Rank Correlation Coefficient is calculated between one word's context and another's. Spearman's Rank Correlation Coefficient is a non-parametric measure of the association between two variables x and y, when the distribution of x or y (or both) cannot reliably be assumed to be normal.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

where d_i is the difference between the rank of the ith x value and the rank if the ith y value and where the n values of x and y have been arranged in ascending order. Finch and Chater have also begun to apply some ideas from self-organising neural networks (Kohonen's work) to their own statistical bigram model, with slightly less successful results. Once a set of word classes has been induced,they argue, these classes can be used to induce grammatical rules, which in turn can be used to improve the original classification.

Brill *et al.* [17] report their attempts to discover the word classes of a language. They use a distributional analysis based on word co-occurrences to cluster classes of words. They describe the requirements for two words to belong to the same word class in set-theoretic terms. Two words, x and y belong to the same class if and only if word y contains all of the features of word x and word x contains all of the same features of word y. The features of x are operationally described in terms of the set of bigrams where x is one of the words.

One of the most significant and prescient works on automatic word classification in the 1970's was presented by Kiss [76]. Here, he develops a psycholinguistically informed computational model of lexical acquisition based on a hybrid Markov model and a proto-connectionist model. He uses a variant of the Canberra measure to describe the dissimilarity between a small sample of words taken from a corpus of mother and child interractions. His classification system also allows for the multimodal nature of words and parallels some results found in child language acquisition research. He too uses traditional statistical techniques to transform his implicit word classification representation into tree-like structures.

The variety of approaches in automatic word classification makes evaluation and comparison of systems difficult. However, McMahon [91] contains experimental evaluations for the systems of Elman, Finch, Hughes, McMahon and Brown. We can make some general comments here about the systems. The phenomenon of ambiguity in linguistics corresponds to a common word classification problem of having to classify words which have multimodal distributions. Many of the the classification systems mentioned above are non-dynamic and they force words to occupy a single place in the classification taxonomy. The systems of Kiss and Pereira, however, both allow for this multimodal property of words by making the classification of words probabilistic. However, Periera's system, like Elman's, Brill's and Kiss's has so far only been able to classify small numbers of words. Schütze's system seems to be able to handle vocabularies of magnitude 10^4. The systems of Brown, McMahon, Finch and Hughes can handle vocabularies of magnitude 10^3, though McMahon and Brown have also developed hybrid systems which can process vocabularies of magnitude 10^4 and 10^5 respectively. Most of the systems require separate clustering to gain access to word classes, but McMahon's representation allows immediate access to word classes. This feature offers practical benefits rather than cognitive plausibility. The systems of Pereira and McMahon cluster in a top-down way, whereas the remaining systems cluster words from the bottom up. Kneser *et al.*'s clustering system requires that all words fit into M pre-determined clusters, plus their classifications are not hierarchical.

2.2 Part of Speech Tagging

In this section, we will describe some representative work in automatic part of speech tagging. The techniques developed here can also be applied to the problem of word sense disambiguation [51], provided that a useful semantic tagging for words exists, and to morphological disambiguation [14]. In all of the work reported below, very high tagging accuracy is reported, though we suggest that the overall percentage accuracy is not the best way to measure part of speech tagging systems. Zipf's law [136] predicts that a small fraction of very frequent lexical items accounts for a very high percentage of the tokens in a corpus; in English, for example, the 1000 most frequent words account for 85% of the tokens. Words like ⟨`the`⟩ and ⟨`a`⟩ are almost always determiners and so even a contextless tagger should score well. Perhaps a better measure of the quality of a tagger would be a measure of how many full sentences it tags with zero errors, one error, *etc.* Elworthy [42] suggests another measure, discussed later.

Church [32] presents an automatic part of speech tagger which uses a dynamic programming technique to maximise the probability

$$P(g(w_i)|g(w_{i+1}), g(w_{i+2})) \times P(g(w)|w)$$

where $g(w)$ is the part of speech tag for word w. The second element in the above expression estimates the prior likelihood that word w has as its part of speech $g(w)$. The first element corresponds to a second order Markov model for parts of speech. He reports 95% to 99% accuracy on a test set when the model is trained on a tagged version of the Brown corpus. Brill and Marcus [18] also describe a system which tags words automatically, but their system does not rely so heavily on pre-tagged corpora. They develop a method of estimating the single most likely tag for a word type, with some help from an informant, and use this to create a prototype word token tagger which achieves 84% accuracy. They then use a small set of tagged utterances as input to a system which learns some word-tag context sensitive transformation rules. The combined system elicits 94% accuracy. Finally, a morphological system tags infrequent words with accuracy 79.5%; since 22% of the word tokens are infrequent, they quote a final tagging accuracy of 90.7% for their test corpus. The tag scheme used in this system is a reduced version (7 word categories only) of the tagged brown corpus.

Kupiec [83] describes an automatic part of speech tagger based on hidden Markov modelling principles. Like the systems of Brill *et al.* and Church, the tags themselves are predetermined linguistic parts of speech. Kupiec reports that the system achieves 96% accuracy. However, unlike the other systems, which use tagged training corpora and hence can directly estimate prior and context probabilities, Kupiec's system is trained on untagged corpora and uses the forward-backward algorithm to estimate parameters. The two sets of parameters for a hidden Markov model are transition probabilities and output probabilities; in part of speech applications, the former correspond to likelihoods of tags at particular points given recently observed context tags, and the latter correspond to likelihoods of particular words given tags. In this system, words are mapped onto equivalence classes in order to alleviate the sparse data problem — this means that, when the hidden Markov model is trained and when it is used to tag new sentences, the first stage involves transforming the words into a stream of equivalence class elements (of which there are 202 in total). These elements correspond approximately to contextless tags: for example, ⟨`information`⟩ might be mapped to 'noun', ⟨`table`⟩ to 'noun-or-verb'. Instead of using first or second order models (bigram and trigram models respectively), Kupiec attempts to model higher order context by adding onto a first-order model some extra network information derived from a (manual) linguistic analysis of the common errors of a simple first-order model. This addition improved tagging accuracy by a fraction of one percent.

The importance of the word-to-equivalence class mapping — functioning as a prior word classification approximation — is addressed by Elworthy [42]. Brill *et al.* [18] already note that a simple system which tags words in a context independent way, by selecting the most frequent tag for each word, can achieve approximately 90% success. If the initial equivalence classification is manually constructed to code for certain linguistic categories, then it is unclear just how much benefit is derived from the non-automatic classification; however, even if it turns out to be a significant factor we might prefer Kupiec's system because the manually constructed equivalence mapping is still contextless and involves less effort than context-based tagging. Elworthy also recognises that overall tag precentages can be misleading and suggest that measuring the percentage of ambiguous words correctly tagged allows a more useful evaluation. He shows also that training from a tagged corpus always leads to better models than starting from equiprobable transition and lexical probabilities. He concludes that, for Baum-Welch re-estimation to be useful, some biasing, either in the transition probabilities, or in the lexical probabilities, must be present in the system before performing the re-estimation. Furthermore, Elworthy identifies three types of re-estimation schedule: classical, early and initial. The classical schedule leads to convergence to an optimum over many iterations; initial scheduling displays a gradual decrease in performance from an initially high state to a sub-optimal convergence point; early scheduling peaks after a small number of iterations, and continues to decline towards a sub-optimal convergence point.

A more traditional automatic part of speech tagger is reported by Hull [65]; a tagged corpus provides direct estimation of tag transition probabilities and lexical probabilities (also called prior probabilities or confusion probabilities). Also, Brill [16] has recently developed transformation tagging — a rule-based automatic part of speech tagging system the rules of which can be learned using an error-driven learning paradigm. He reports results which are better than stochastic Markov models — 96.5% accuracy — and claims further that his method is better because it can capture linguistic information directly, though he offers no arguments to support the claim that explicitly rule-based systems are better than stochastic systems. Certainly, the rule based system is more informationally compact than an equivalent Markov system. During training, the output from a approximate tagging system is compared with the manually tagged equivalent and rules are induced which reduce the error (difference).

2.3 Segmentation

Sentence structure can be described in terms of constituents. Grammar determines the rules of combination (syntagmatic rules) of these constituents. In any segment of text, there are many ways of sub-dividing that text. From a language processing perspective, the most interesting way is by dividing it into the constituents of the grammar which is said to generate that language. Constituents can be hierarchical — at one level, the text is divided into sentences, at another, into sub-sentential constituents.

Syntactic analysis can be considered to consist in segmenting and interpreting sentences. Of course, like many aspects of natural language processing in humans, there is much feedback between these systems; however it may still be useful to examine each process separately. The second process, interpretation, involves making judgements (often linguistic) about the various elements of a language stream; the previous section summarised some attempts to analyse the language stream through automatic tagging. In this section, we introduce some research in the area of automatic segmentation. The aim of many of the researchers introduced in this and the previous section is to advance components of an overall system which can parse language streams: here, parsing is segmentation plus tagging, where

tagging is applied not only to words, but to higher level constituents. Segmentation itself can be flat or hierarchical: a flat segmentation takes a stream and identifies significant boundaries, whereas hierarchical segmentation produces a bracketing which allows for the construction of a parse tree for some given linguistic input stream. Often, similar techniques are applied to both types of segmentation, although hierarchical segmentation is more difficult than flat segmentation.

Faulk *et al.* [44] make the convincing claim that successful human-machine communication using natural language must be preceded by an account of language acquisition which is stochastic and which explains how successful grammatical competence is reached through exposure to a possibly degenerate and certainly finite language sample. Their approach is structural linguistic in the sense that they assert that there is enough structure *in* language for a learner (human or machine) to induce that structure. They describe a flat segmentation system which operates on letters, though they claim that it can operate on phonemes and other linguistic levels. They use a measure called the variety index. Variety index minima correspond to constituent ends and hence can be used to segment the linguistic stream. The index is calculated as the product of the mean of the proper left and right bi-string ratios, which themselves are the significant conditional maximum likelihood probabilities when processing in a leftwards and a rightwards direction. That is, if the letter-stream $\langle$`thecatsatonthemat`$\rangle$ is being processed and we are interested in the variety index of the sixth letter (the $\langle$`t`$\rangle$ of 'cat'), then left bi-string ratios might include $P(\langle\texttt{at}\rangle|\langle\texttt{a}\rangle)$, $P(\langle\texttt{cat}\rangle|\langle\texttt{ca}\rangle)$ and $P(\langle\texttt{ecat}\rangle|\langle\texttt{eca}\rangle)$, as long as non-zero values exist for the n-gram probability estimates. Similarly, some right bi-string ratios might include $P(\langle\texttt{ts}\rangle|\langle\texttt{s}\rangle)$, $P(\langle\texttt{tsa}\rangle|\langle\texttt{sa}\rangle)$ and $P(\langle\texttt{tsat}\rangle|\langle\texttt{sat}\rangle)$, with the same restrictions. Left and right bi-strings are averaged of give a single pair of left and right bi-string values. The product of these values gives us the variety index at the sixth position in the input stream. In effect, the bi-string ratios estimate a very simple version of a weighted average language model [103]. One advantage of segmenting a given input is that left and right context can be exploited, whereas language models destined for speech recognition tend to limit themselves to left contexts only. Faulk *et al.* limit themselves to using corpora which are produced by simple, modestly-sized finite state grammars but their results indicate a degree of success.

Brill *et al.* [17] and Church [32] develop the use of mutual information within computational linguistics in several interesting ways. They apply the concepts to grammar induction and automatic hierarchical segmentation. They use a corpus which has been annotated with parts of speech and develop a constituent boundary segmenting algorithm which takes a sentence from the Brown corpus, such as

```
He directed the cortege of autos to the dunes near Santa Monica
```

and segments it as follows:

```
(He(directed((the cortege)(of autos)))((to(the dunes))(near Santa Monica)))
```

The Brill *et al.* algorithm uses mutual information *minima* to discover constituent boundaries; the same insight has been applied in the work of Wolff, using a rule-based system [132, 133], by Elman [40, 41], Pollack [108], Jordan [71], and Gasser [52], using connectionist architectures and, as we have just seen, by Faulk *et al.* [44], extending Harris' idea of a *variety index* value being minimised at constituent boundaries It is also related to the idea of discovering useful schemata with genetic algorithms [55, 61]. The same periodic rise and fall in uncertainty is described using information theoretic terminology by Shannon [124]. Also, Attneave [3] describes a similar phenomenon in a model of visual perception — information is highest along contours and boundaries in a visual image, and highest of all when the rate of change of the boundary is highest.

In Brill *et al.*'s formulation, word class n-grams derived from a tagged corpus are examined in order to find likely constituent boundaries, or *distituents*. Their hypothesis states that a form of mutual information called *generalised mutual information* will be able to identify distituents. This can be explained further by an example.

If the class n-gram is ⟨`det noun verb`⟩ and we are looking for the most likely constituent boundary — *e.g.* let us assume that the best partial segmentation is `((det noun) verb)` — then the probability $P($⟨`det noun`⟩$)$ should be significantly higher than $P($⟨`det noun verb`⟩$)$. Informally, this captures the intuition that good constituents should occur in many contexts. In terms of the task of predicting which class comes next, the entropy should remain low, and possibly even get lower, until the constituent ends, at which point the entropy for the next class should be significantly higher. These high entropy break-points in effect mark off the structure of particular utterances.

This method leads to successful (unlabeled) parse tree estimations for a test set of unconstrained free discourse; for sentences of length less than fifteen, the parser averages two errors per sentence, rising to between five and six errors for sentences between sixteen and thirty words long.

Recurrent connectionist architectures [41, 40, 71] have been applied to the task of discovering the structure of language from its serial expression. In these cases simplified grammars are used, with restricted vocabularies. Connectionist researchers also recognise that high entropy break points represent useful ways to proceed with the discovery of linguistic structure. Elman [40] uses slightly different terminology — he notices that time-varying error signals in recurrent nets can provide clues to structure.

Reilly [114, 115] develops this work by training another neural network to take as input the hidden layer activation state and to output a partial but explicit parse. Thus, in his system, just as in Brill *et al.*, algorithms exist which can automatically parse incoming streams of words, using the structure which is implicit in language, but without having an explicit traditional linguistic component. Brown *et al.*'s class-based n-gram language model can also be described as a stochastic grammar which is implicit in the frequency statistics, and also distributed.

Scholtes [120] implements his *Data Oriented Parsing* system with a Kohonen feature map. The system uses structural features together with statistical information from a corpus. His parser produces a set of ranked parses for an ambiguous sample sentence; it produces partial parses for incomplete sentences, wrong sentences and new sentences which contain several totally unseen words or structures.

Recently, Juola *et al.* [74] have applied a similar technique to the discovery of morpheme boundaries in words. They develop a system based on the assumption that entropy rises at morpheme boundaries. They report 47% accuracy when their system is trained on a 31,000 word corpus (2,500 distinct words), compared to judgements of a native speaker of English. Also, Brent [15] offers a similar system, using the slightly different terminology of minimal generative explanation, for discovering morphemes. He claims that the idea can be extended to syntactic segmentation. In one experiment, based on frequency statistics from the 8,000 most frequent words in a corpus, Brent reports that 79% of the morphemes his system discovered corresponded to attested morphemes.

Also, Kozima *et al.* [80] extend the idea of language stream segmentation to higher levels by using a semantic net (constructed from the LDOCE dictionary) to provide data on the semantic cohesion between words and by estimating the lexical cohesion profile of a text. By this method, text streams can be segmented into coherent semantic 'scenes'. This is a structuralist approach to automatic text (proto-)understanding. Again, a similar rise-and-fall pattern can be observed when moving from one coherent scene to another.

2.4 Language Models for Recognition

Statistical Language modelling for speech recognition applications has often progressed using the least cognitively plausible models of language [28, 111]. However, it has led the way in terms of the development of sophisticated probability estimates. Not only were statistical language modellers the first to produce robust large vocabulary language model systems by incorporating probabilities of linguistic events (in the first and simplest case, linguistic events being equated with occurrences of certain words), but they were among the first to identify the weaknesses in the models and to suggest involved statistical and information theoretical advances on these models. Nor have many statistical language modellers been blind to the lack of cognitive plausibility of their models, such considerations being, in many cases, peripheral. Despite this sometimes staunch engineering approach, recent statistical language modelling systems have shown signs of moving beyond the traditional finite-state word based grammars which underpin much work in this area.

Broadly speaking (and mainly for the purpose of exposition), we can say that there are three areas of investigation which statistical language modellers are currently examining: word distribution modelling, word context modelling and language model integration. Research often overlaps all three areas. Firstly, and perhaps most importantly, work has continued in advancing improved probability models of rare events. Many word types occur very infrequently or not at all in even the largest training corpus; also, when a system needs to estimate word bigram, trigram and higher order n-gram probabilities, the problem of accurate probability estimation becomes even more acute. A second area of statistical language modelling research investigates more sophisticated models of context. Researchers in this area need to estimate the probability of some event given the occurrence of some linguistic context. This context can be as simple as the identity of the previous word in the word stream, or it can be as complex as a binary decision tree which allows for higher level linguistic knowledge. One specialised subset of these models, which can be constructed quickly, is the class-based language model. Some work in this second area began not as an attempt to instill some cognitive plausibility in these models, but as a practical engineering response to the difficulty of modelling the probabilities of rare words. Finally, most statistical language modelling systems maintain several models of the distributions of word probabilities. Again, this is an engineering response to the sparse data problem. Some work has continued in designing more efficient and more successful ways of combining these sources. Eventually, the improvements made by statistical language modellers will filter through to the rest of the language processing community, who should incorporate these superior probability models into their part-of-speech tagging, word classification, grammar induction, sentence parsing and machine translation systems.

Suppose we want to estimate the probability of some word w in a stream of words (corpus). We can consider the stream, length n, as the set of results of an experiment, repeated n times, the outcome of each trial being partitioned into two mutually exclusive and exhaustive event sets, 'w occurred' and 'w did not occur'. By describing a corpus as the result of such an experiment, we have assumed that the probability of a word is independent of previous occurrences of other words; this is obviously false. The probabilities of words are dependent on the previous words in the stream both syntactically and pragmatically: if the previous word is ⟨`the`⟩ then one is much less likely to observe the word ⟨`of`⟩ next; if ⟨`apple`⟩ has occurred recently, then it is more likely to occur again soon, based on the assumption that the topic of conversation involves apples. However, when we make the assumption, we have a situation which can be modelled by the binomial distribution.

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \tag{6}$$

where p is the probability 'w occurred'.

The expectation $E[X]$ can be estimated as

$$E[X] = np \tag{7}$$

which allows us to estimate p *a posteriori* from the corpus. This estimate is called the maximum likelihood [6] since it estimates a value for p which makes the observed results (the corpus stream) maximally likely. Whenever the variance, $np(1-p)$ is sufficiently large, the distribution becomes approximately normal. This assumption has allowed researchers to develop some robust statistical language models, but there are still many ways to make improvements.

Dunning [38], for example, points out that the majority of word types are so rare that statistics which rely on normality in the underlying distributions become less accurate. Gale *et al.* suggest that for a bigram language model the sparse data problem gets worse with larger training data — that is, $V > O(\sqrt{N})$, where V is vocabulary size and N is corpus size. Sampson [119] makes a similar point about the variety of noun phrases in a corpus. If this is so, then larger corpora will not solve the sparse data problem (though they will lead to systems which perform better). With rare events, the normal approximation tends to over-estimate the significance of the event. The degree of error varies, depending on the rarity of the event. For example, a mutual information estimate between the two word events (see equation 1) ⟨`the`⟩ and ⟨`carburetor`⟩ involves an error which is significantly different from the error between ⟨`the`⟩ and ⟨`car`⟩. This is important, for example, if you want to use these values to automatically classify words [92]. Fisher *et al.* [50] apply the t-test to resolve relative pronoun attachments, using a small (500,000 words) corpus. However, they do not rely on lexical frequencies, the majority of which are distributed so sparsely that the normal assumption obviously does not apply; instead they use semantic features of words (from a feature vocabulary of size 67). In this case, the normal assumption is violated less severely. Of course, this approach relies upon a corpus which contains words which are tagged with semantic features.

Dunning proposes using the likelihood ratio test, which is applicable in cases where we cannot safely assume that a word is distributed normally. First, he defines a likelihood function

$$H(\omega; K) \tag{8}$$

which gives the probability of observing experimental results K for a fixed model (*e.g* binomial) of parameter set ω. For the binomial case, there is only one parameter, p, the probability of a successful outcome, and two observations can be made for each set of experiments: the outcome occurs k times in n experiments. The likelihood ratio, λ for some hypothesis which corrsponds to parameter subspace Ω_0 is

$$\frac{\max_{\omega \in \Omega_0} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)} \tag{9}$$

For example, we could test a hypothesis about the strength of the collocation between two words, A and B. The null hypothesis can be stated as:

$$H_0 : P(A|B) = P(A) \tag{10}$$

With observations k_1 and n_1 corresponding to the number of times A is seen after B and the number of times B is seen, respectively, and k_2 and n_2 corresponding to the number of times A is seen in the corpus and the number of words in the corpus, respectively, then

$$\lambda = \frac{\max_p H(p, p; k_1, k_2, n_1, n_2)}{\max_{p_1, p_2} H(p_1, p_2; k_1, k_2, n_1, n_2)} \tag{11}$$

The parameter p is maximised when it equals $\frac{k_1+k_2}{n_1+n_2}$, p_2 is maximised when set to $\frac{k_1}{n_1}$ and p_2 is maximised when set to $\frac{k_2}{n_2}$. Dunning reports some preliminary success using this test to find collocations.

Badalamente *et al.* [4] report that the distribution of new words in a word stream (their corpus consisted of famous poems) can be modelled as a Poisson process. The more time which has elapsed since the last new word, the higher the probability that the next word will be new. They also claim that different model parameters identify different authors. A Poisson process, parameter μ is defined by the following p.d.f. for the discrete random variable X:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \tag{12}$$

for $x = 0, 1, \ldots$. The distribution can be used as an approximation of the binomial distribution with large n and small p ($\mu = np$).

Katz [75] uses a formula by Good [57] (with Turing) to improve upon the maximum likelihood estimate for rare events. If an event occurred r times in an n-event sequence, then the maximum likelihood estimate is r/n; in the Turing-Good model, the probability of such an event is $r*/n$,

$$r* = (r+1)\frac{n_{r+1}}{n_r} \tag{13}$$

where n_r is the number of events which occurred r times in the corpus. For example, if the events in question are occurrences of bigrams, in a system where the unigram model is fixed (vocabulary known), then we can estimate a non-zero probability for bigrams which did not occur in the corpus ($r = 0$) as $\frac{n_1}{n_0}$; n_1 is the number of bigrams which occur once in the corpus and n_0 is the number which do not occur at all ($V \times V - \sum_{r>0} N_k$). This estimate requires only that the event be modelled binomially. The advantage is that no event (*e.g.* bigram) is assigned a zero-probability estimate); the system is easy to implement and leads to measurably better language models.

Church *et al.* [30] describe an empirical version of the Turing-Good estimate, called the held-out estimate, attributed to Jelinek. The Turing-Good estimate can be considered to be the expected frequency, in a new corpus, of a bigram which occurred r times in the original corpus. In the held-out method, N_r is the count of all bigrams which occur r times in the first corpus and C_r is the total count of those bigrams in the second (held-out) corpus. The heldout estimate $r*$ is calculated as C_r/N_r. A version of this model is known as the deleted estimate. Here, held-out estimates are calculated and then the two corpus roles are reversed: the held-out corpus becomes the retained corpus, and vice-versa. The final estimate is calculated as:

$$r* = \frac{(C_r^{01} + C_r^{10})}{(N_r^0 + N_r^1)} \tag{14}$$

where N_r^0 is the number of bigrams in training corpus 0 and C_r^{01} is the total number of occurrences in training corpus 1.

One limitation with these methods is that they assign the same probability to all events which occur zero times (or once, twice, and so on). Church *et al.* suggest a way of allowing us to partition the set of bigrams which occur r times, by using a second source of information — component unigram frequencies. Although words clearly are not generated independently, $P(x)P(y)$ is still an informative source of information about the zero-frequency bigram $\langle$xy$\rangle$. These enhanced models lead to qualitatively better results (enhanced Turing-Good is better than enhanced deleted estimate).

Nicholis *et al.* [100] present a radically different and highly original model of linguistic structure. In their model, a chaotic map is partitioned into an abstract alphabet; dynamically iterating this map

produces a text. When some collective parameters of this text are examined — Zipfian statistics, mutual information, and Markovian profile — the results are similar to those produced by a natural language. Chaotic maps can account for certain basic syntactic requirements: context history, selectivity of few keywords and polarity (*i.e.* sentences are not the same when read backwards). They use the logistic map as their generator:

$$x_{n+1} = Cx_n(1 - x_n) \tag{15}$$

with various values of parameters, $C = 3.57, 3.7, 4$. The correlations between the artificial and natural texts (with respect to several macroparameters) suggests a similar underlying chaotic generator.

The final example of work in the area of improving basic models of word distribution is by Kuhn *et al.* [82]; they develop a cache-based system which attempts to model the non-stationary aspects of word distributions: natural language tends to be about particular subjects, at a local level. A simple way to model this is to exploit the fact that once a word has appeared in a text stream, the probability that it will occur at some point in the immediate future is higher than the global *a priori* (unigram) probability predicts. In their system, words are associated with parts of speech. The probability component $P(w|g)$, estimating the likelihood of word w given class g is normally calculated globally (see equation 4); however, each part of speech is also associated with a cache which stores the n most recent examples of words which are parts of speech of that type. Thus there are two ways of calculating $P(w|g)$; these models are weighted together, leading to a combined model which supports sensitivity to local word frequency distributions whilst maintaining a global, reliable, but less informative model. The authors claim a threefold reduction in perplexity when the model is tested.

The second strand of statistical language modelling involves the estimation of better contexts. Early work in the area considered the most recent n words [66, 67] as context — $P(w_{n+1}|w_1^n)$. Word-based contexts then came to be seen as simple partitions of a general context $\phi(\sigma)$ — $P(w_{n+1}|\phi(\sigma))$. It became clear that there were many functions ϕ which partitioned the context, σ. One popular and easy to implement context function mapped contexts together if and only if each of the most recent n words corresponded to certain parts of speech — $P(w_{n+1}|g_1^n)$. These class-based language models [68, 20, 37] improved performance (as measured by perplexity); however, whilst the models are significant advances on purely word-based ones, their cognitive plausibility is still questionable. Class-based language models usually require previously tagged training data, with tags usually corresponding to familiar linguistic tags (see [37, 82]); however, some systems automatically generate their own (usually contextless) tags from raw text (see [20, 92]). McMahon *et al.* [92] also describe a multi-level class-based language model which allows the system to fall back on reliable but relatively uninformative class-based statistics where instances of a word stream are rare, but to exploit more fine-detailed class (and word) information where it exists.

Bahl *et al.* [5] suggest an even more interesting theoretical description within which the idea of contexts can be situated. They argue that the conflation of contexts into equivalence classes which only deal with the previous n occurrences — *i.e.* n-gram language models — is useful but unnecessarily restrictive. For them, the design of a context structure should allow it to contain information about words at an arbitrary distance from the current prediction position, provided that this information is statistically reliable. Recent-history equivalence classes based upon word information only allow for questions of the type : 'was the last word *the*?' or 'was the second-last word *cat*?' With their context structure Bahl *et. al.* describe the space of possible binary decision trees and suggest methods of discovering some of the more useful ones. They construct trees which minimise the average entropy of the leaf distributions and construct a hybrid trigram-tree language model which results in a perplexity 13% lower than the pure trigram model, and 9% lower than a pure tree model.

The final area of activity within the field of statistical language modelling involves finding efficient ways of combining language models. A hugely influential way of combining models is by considering them to be Markov sources [66]. Often, the component language models have complementary strengths and the hybrid is constructed in such a way that maximum weight is given at any one time to the most informative and reliable component. The greater the degree of flexibility in distributing weights across the components, the higher the performance. In most cases, combinations of unigram, bigram and trigram models are weighted together. For example, the simplest interpolated trigram language model contains two independent weights, λ_u, λ_b — the trigram weight, $\lambda_t = 1 - (\lambda_u + \lambda_b)$. Given this hybrid, some method needs to be introduced which can optimise the independent weights and hence optimise the performance of the hybrid.

A more complicated hybrid system can be built by making the weights depend on some other, easily calculated parameter. For example, in the estimation of test-set perplexity, the frequency of the previously processed word is readily available. Whereas in the simple hybrid system, each stage of the processing gives fixed weight to each of the three components, in this new hybrid, the weights can vary throughout the test set. The problem of selecting a set of optimal weights becomes more apparent in this case, since there are usually hundreds of different frequency-dependent weights.

In order to use a parameter estimation technique from Markov modelling theory [110, 103] it is useful to think of the hybrid language model as a Markov chain with two types of arc — emitting and non-emitting. Figure 1 shows a single transition for the simple hybrid case. The probability of word

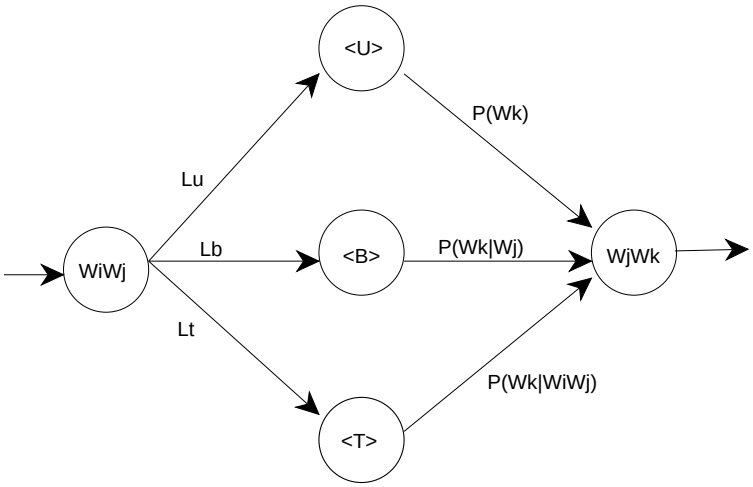

Figure 1: Section of a Markov Chain showing the transition from the state corresponding to word-pair w_i,w_j to the state corresponding to word-pair w_j,w_k. The first three arcs, Lu, Lb and Lt correspond to the non-emitting unigram, bigram and trigram transition weights λ_u, λ_b and λ_t. The second set of arcs correspond to the maximum likelihood conditional probabilities of the word w_k, for unigram, bigram and trigram language models.

w_k following the segment $\langle w_i, w_j \rangle$ is equivalent to the two-stage transition from the state $\langle w_i, w_j \rangle$ to the state $\langle w_j, w_k \rangle$, which is equal to the sum of all ways of making that transition; that is

$$P(w_k) = \lambda_u \times P(w_k) + \lambda_b \times P(w_k|w_j) + \lambda_t \times P(w_k|w_i, w_j)$$

The emitting transition probabilities are estimated as usual in a maximum likelihood way, using a training text. The non-emitting, weight probabilities are estimated using a separate unseen text.

A simplified version of the *Forward-Backward* algorithm [103] can be used iteratively to optimise a set of initial parameter values, to an arbitrary degree of significance. The update equation for the jth weight, λ_j out of L language models is as follows :

$$\lambda'_j = \sum_{i=1}^{n} \frac{\lambda_j \times P_{LM_j}(w_i)}{\sum_{k=1}^{L} \lambda_k \times P_{LM_k}(w_i)} \tag{16}$$

where the held out corpus is n words long and $P_{LM_j}(w_i)$ is the probability estimate of word w_i, using the jth language model. This procedure has been shown to lead to Markovian language models where $P^t(w_1^n) \leq P^{t+1}(w_1^n)$ — *i.e.*, given that the held out text is a sufficiently representative sample of the language being modelled, then the simplified Forward-Backward algorithm makes the held-out text iteratively more likely. The held-out text should be disjoint from the test and training sets to prevent over-learning of those texts.

With the more complicated interpolated language model — where lambda values depend on frequencies, $\lambda'_j(f)$ is calculated using an equation similar to 16 above, except that only those words w_i are used which come after a word whose frequency is f.

Interpolating language model components requires that the training data is fixed; new data requires the system to be re-trained. Recently, some work has begun on finding mathematical relationships between language model components [101]. O'Boyle *et al* [103, 102] have designed weighted average language models. These models also have the advantage of being able to exploit higher order n-grams which occur in statistically significant numbers in a corpus, without leading to the explosion of training parameters which a similar extension would entail with interpolated language modelling. This model is described as follows :

$$P(w_k|w_1^{k-1}) = \frac{\sum_{i=1}^{m} \lambda_i \times P_{ML}(w_k|w_{k-i}^{k-1}) + \lambda_0 \times P_{ML}(w_k)}{\sum_{i=0}^{m} \lambda_i} \tag{17}$$

where there are informationally 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 normaliser. It has been found that

$$\lambda_i = 2^{|w_{k-i}^{k-1}|} \times \log f(w_{k-i}^{k-1}) \tag{18}$$

where $|w_{k-i}^{k-1}|$ is the length (in words) of the segment, results in a useful language model. This approach might prove a more adequate platform for implementing a multi-level class-based language model [92].

2.5 Grammar Induction

One linguistically well supported method of modelling language is by specifying a grammar. The quality of a particular grammar can be assessed by measuring its performance at parsing (assigning hierarchical bracketing and appropriate labels for constituents at all levels) some test set of sentences. Many of the techniques which we have examined so far have been designed precisely to avoid some of the difficulties in constructing broad coverage traditional grammars. However, computational linguists who work on grammar induction have also made some significant recent contributions to natural language processing. Many have done so by extending the structure of their grammars to include a stochastic element. This strain of computational linguistics therefore, retains many close connections with mainstream linguistics and cognitive science, yet is beginning to produce robust parsers to rival systems often (wrongly) considered to be purely statistical and grammar-free.

Before we begin our brief review, we recall five types of grammar, from most powerful to weakest, which Chomsky considered worth discussing as models of natural language; they are: recursively enumerable, recursive, context sensitive, context free and finite-state. Chomsky has spent some energy arguing against variants of finite-state grammars; however, as we have already seen, much work on natural language modelling has already been carried out with stochastic versions of finite state grammars (for example, word-based n-gram language models). This has been the case for at least two main reasons, one pragmatic and one theoretical: firstly, we note that stochastic finitary models are less complex than their more powerful rivals and can be built with less effort; we also note that the adoption of such grammars allows computational linguists to expliot the mathematical techniques of equivalent systems — namely Markov models; secondly, computational linguists may prefer finitary grammars for theoretical reasons: perhaps they consider them more compatable to a holistic (anti-modular), empiricist (anti-rationalist) model of cognition. The former reason is more resistant to criticism than the latter.

Grammar induction researchers model natural language with phrase structure grammars (context free and sensitive) or with grammars which are equivalent to phrase structure grammars. Since n-gram word-based models are stochastic and finitary, researchers who build them are also inducing grammars. The next most powerful grammar is the context-free grammar, which recently has become a popular model base onto which researchers can add probabilistic weights.

Krotov *et al.* [81] use a corpus which contains some diagrammed sentences and some tagged sentences, from which context-free rules and associated probabilities can be extracted. This is a significantly different approach from one which attemps to derive a grammar from a raw corpus of words, or even from a corpus of part-of-speech tags. Krotov *et al.* describe their method as 'grammar extraction' and contrast it to 'grammar training'. Since grammar extraction attempts to induce stochastic grammars from many instances and grammar training uses no parse instances, it is clear that grammar training is a more difficult task than extraction. Krotov *et al.* found that the number of rules extracted from the corpus and the number of words in the corpus display a log-log root relationship. Their conclusion — that it is still not clear if the number of rules will have a bounded upper limit — is consistent with a finding of Sampson [119] which he uses to argue against the grammaticality/ungrammaticality distinction in natural language. Not only is the stochastic grammar extracted by Krotov *et al.* poor in terms of coverage, but the set of extracted rules makes heavy computational demands; this is such a problem that, instead of attempting to deal with lack of coverage, they suggest ways of eliminating many of the existing rules in the rule-base (for example, by means of deleting many improbable rules, or by unifying similar rules).

Resnik [116] argues that stochastic context-free grammars are not good models of natural language because they are not particularly sensitive to lexical context. Instead he suggests that probabilistic tree-adjoining grammars offer a more promising framework. Tree-adjoining grammars are generalisations of context-free grammars — the standard 'substitution' rule of context-free grammars is supplemented by an 'adjunction' rule. Lexicalised tree-adjoining grammars consist of tree-fragments which have a lexical item 'anchoring' the tree. Like Krotov *et al.*, Resnik offers no immediately practical way to induce the grammar, though he suggests that the technique of parameter re-estimation using the Inside-Outside algorithm could estimate the probabilities of the model.

Carroll *et al.* [24] report some research aimed at learning probabilistic dependency grammars using a part-of-speech tagged corpus as training data. Given this approach, they state a preference for utility over theoretical purity in their work. The also state an interest in modelling some diachronic elements of language by allowing the learning mechanism to constantly adapt to changes in language use. Adaptable grammars also have obvious pragmatic value. In their attempt to learn grammars from

artificially generated languages, they arrange the training corpus in such a way that short sentences are presented to the learning algorithm before long ones. A similar idea has been suggested by Elman [41]. For each sentence which cannot be parsed by the current grammar, rules are added which allow a successful parse. The number of new rules for any sentence is guaranteed to be finite with dependency grammars. The next stage involves using the inside-outside algorithm to estimate probabilities for these rules. It is unclear whether learning takes place if a sentence can be parsed by the grammar in its current state: if it does not, then an opportunity for learning has been lost since the hypothesised parse suggested by the grammar might be unlikely, yet no alternative parses would be tried. If learning does take place then the algorithm will need to re-train its probabilities for the entire grammar after every new sentence. The authors claim that their incremental approach allows them to start with an empty grammar and to evaluate a less constrained rule-space. They add that estimating initial rule probabilities is a problem for the inside-outside algorithm because its performance is sensitive to these initial conditions. Again, their incremental approach may be less prone to lead to local minima grammar states — for example, a learning experiment was carried out with training sentences being presented in batches as opposed to in series. With random (near uniform) starting probabilities for rules, 300 different experiments resulted in 300 different grammars. Even though dependency grammars are constrained context-free grammars, Carroll *et al.* found that they had to add more constraints to their model to achieve grammar induction, even with artificial training texts. They add rules which determine, for any non-terminal symbol, which non-terminals can be the head of a rule involving that symbol. This effectively prevents the algorithm from learning a structurally useless 'memorising grammar'. The authors identify two problems with their system: first, the rules it induces are overly specific and second, the developing grammar, as it is taking shape, looses its capacity to overcome local minima traps. This second feature is important because the ordering of the training corpus inevitably introduces skews into the grammar from an early stage, from which it will find it impossible to recover. The authors add a component to their learning system which learns from pseudo-negative evidence: at any stage, the grammar can generate sentences and assign probabilities to them; as the difference between the assigned probability and the observed probability of the sentences becomes great, the rules which generated the sentence are re-examined and possibly removed. In practice, only rules which generate high-probability pseudo-sentences are examined, and they are removed only if their observed probability is low.

As we have seen earlier, the preponderance of many very specific rules was also a problem for Krotov *et al.*. One solution both sets of researchers apply is to remove low probability rules by setting a rule threshold. Carroll *et al.* also attempt to reduce the size of a grammar by trying to eliminate low probability rules without reducing coverage. However some sentences occur once in a corpus and require a unique parse (that is, they require low probability but irreducible rules). Carroll *et al.* treat these as exceptions, though Sampson [119] offers us an argument against this practice.

Bod [13] describes a performance based grammar induction system which superficially appears to be based on a stochastic context-free grammar. He considers the best model of the language processing ability in humans to be based on a corpus of all the syntactic and semantic structures a typical human experiences, together with a lexicon. The main difference between this approach and one which uses a stochastic context-free grammar is that Bod replaces the parsing process with a kind of analogy-construction between input sentences and parts of the structured corpus; the goal is to reconstruct a good structural description of the sentence by direct or indirect (through abstraction) comparison with previously witnessed structures. Eventually, a maximally likely 'parse' is selected and this new parse is added to the corpus of linguistic experiences. The maximally likely parse is selected after examining only a sample of the possible sets of constructions, to avoid immense computational overheads. This

system shares some common features with example-based translation (see section 2.7).

Formal language theory provides the theoretical background for grammar induction by syntactic pattern recognition [56] where input vectors are considered to be well-formed expressions of some abstract grammar G_i. Naumann and Schrepp [98] suggest one method for inductive learning of a grammar which will parse a given corpus. They use an incremental learning algorithm which produces a sequence of grammars, each of which parses the corpus more successfully. New sentences from the corpus are parsed to produce partial structural descriptions; a set of new grammars for the corpus is created which will parse the sentences in question, and the grammar which makes the smallest inductive leap is picked to be the new grammar. The main disadvantage with this approach is the danger of over-generalisation. Systems which do not use probabilities to guide selectional preference find modelling language acquisition more difficult: in effect, all grammatical sentences are considered equiprobable.

Work on formal languages as models of natural language follows from work by Chomsky [28], by Gold [54] and more recently work by Berwick [12] on application of the Subset Principle — a technique which, while restricting itself to positive-only input the order of which is constrained, guesses the narrowest possible language compatible with the data given so far. This makes the hypothesis maximally disconfirmable.

Solomon and McGee-Wood [126, 125] have applied categorial grammar [90] to the task of learning a grammar. The process is semi-automatic and uses as its corpus a sample of childrens' utterances.

In categorial grammar, a small number of atomic categories (usually s for a sentence, np for a noun-phrase and n for a noun) are postulated and all other words are defined in terms of complex categories made up of some combination of these primitives. For example, a word which was an intransitive verb might be described by the functional category $s\backslash np$, which indicates that the word in question is of that category of words which, when prefaced by a noun phrase, produces a sentence. Due to the complex nature of the categories, the resulting lexicon implicitly captures the full richness of grammatical relations. In the system, word ambiguity is dealt with by allowing a word to have multiple categories.

Another researcher, Wolff [132, 133] suggests that the bootstrapping problem can be solved through data compression techniques. Like Finch and Chater, Wolff exploits the informational redundancy in a corpus in order to develop a grammar which represents that corpus. He extends this idea by suggesting that the development of language in humans is driven by the minimisation of information storage and retrieval. He defines *efficiency* in a body of information as *power / size*, where power is the expressive power of the body of information — *i.e.* the non-redundant information it contains, and size is the number of bits in the body of information. A grammar, then, which codes for a set of utterances, can be considered to have captured the power of the utterances, but is much smaller in size. This gives him a measure to compare different grammars which cover a given set of utterances.

Genetic Algorithms [55, 61, 89] allow learning through a process of natural selection with respect to an optimisation task. A population of estimates of solution hypotheses for a given problem are compared using a fitness function. There are also mechanisms which supply mutation or crossover, or both; these randomise parts of a hypothesis, generating new hypotheses and preventing the learning from settling into local minima.

In classifier systems (see [62, 7] for introductions to the theory of classifier systems), the rule of transition from instance segment to class segment can be seen as a simple if-then rule, which is the atomic element of a classifier system. Those if-then rules which tend to cover the test segments in the most efficient way are rewarded; unhelpful or wasteful rules receive less reward and tend to die out. When linked to a genetic algorithm new grammatical sub-strategies are introduced into the fray. The

system then tends to evolve towards better approximations of the underlying grammar of the tested language. This approach is like an evolutionary version of Wolff's rule-based symbolic system and Schrepp's more formal symbolic system.

Antonisse [1] develops a reformulation of genetic algorithms so that they can represent any problem which can be described as a formal grammar. He does this by re-defining the crossover operator so that the newly created string is well-formed in some grammar. This is achieved by tagging the trailing and leading edge of each split string with a tag which captures the relevant string fragment's position in the phrase structure from which it originally came. Now, two strings can link if and only if their trailing and leading tags can be unified. Koza [79] has also developed work along this direction in his genetic programming paradigm, though Antonisse claims that his own work subsumes Koza's. This development in genetic algorithm theory is equivalent to a move away from n-gram finitary models of language (which could only search for structure in surface strings) towards more complex (*e.g.* phrase structure) grammars.

Wyard *et al.* [134] have designed a single layer higher order neural net which takes as input tuples of word class units — *e.g.* ⟨`adj noun`⟩ — and tries to identify those tuples which help in grammaticality decisions. The input sentences are either grammatical or not, and the net is trained, using a form of punishment learning, to output a simple binary grammaticality decision. This system could be used as a pre-parse filter. Positive sentence samples are generated from a context free grammar and negative samples are randomly generated. The neural net was also supplied with extra information which made sentence boundary determination trivial. This contrasts with Elman, whose net has as input a constant stream of words with no explicit sentence boundaries.

2.6 Modelling Specific Linguistic Phenomena

Researchers who use training corpora which contain part-of-speech tags, phrase structure or semantic tags can also attempt to model some specific language phenomena: selectional restrictions, prepositional phrase attachment and various kinds of disambiguation.

Resnik [117] brings the techniques of information theory to a noun taxonomy which has been constructed by hand in the form of a semantic network. He claims that this helps in elaborating an empirically adequate theory of selectional constraints, which he bases upon the concept of 'preferred association'. Essentially, information theoretic measures provide each word with a 'selectional profile', which can be compared numerically with other such profiles with respect to particular associations.

The model he constructs seems to deal well with the traditional examples of selectional constraint; it also shows useful properties when a class of verbs is analysed with respect to its argument realisation properties; finally, Resnik reports success when the system is used to syntactically disambiguate lexical items in an unconstrained text.

Resnik's work uses the WordNet lexicon, which is described by Beckwith *et al.* [11]; it has been constructed on some principles of human lexical organisation developed by psycholinguists. It provides a rich representation for language modellers to use, whether they base their models on X-bar theory or on statistical collocations and mutual information. WordNet currently contains information about 64,000 different nouns, verbs and adjectives. It has been developed from a synchronic perspective, rather than the usual diachronic perspective of most dictionaries.

Burger and Connolly [23] provide good examples about the danger of estimating statistics of high-level — that is, non-surface — linguistic phenomena: they construct a Bayesian Network partly by hand (designing its overall structure) and partly from frequency information (estimating the arc transition probabilities). They are forced to calculate corpus-derived statistics of events which are

the constructs of linguists — for example, 'discourse focus', a phenomenon which is unseen in a raw corpus, and which may be constructed only upon acceptance of a particular linguistic theory.

Some researchers prefer to use hybrid statistical and syntactic models to improve performance in disambiguation, recognition, part of speech tagging [18, 32] or generation. Rohlicek, Chow and Roucos [118] have reported success using a small corpus and a set of manually constructed sentence templates, unto which sentences are mapped. Then, using the reduced number of training parameters which their sentence template system allows them, they construct a useful Markov model of language which can automatically tag words. Basili, Pazienza and Velardi [8] also add in some syntactic and semantic information to improve the performance of their system. This work extends statistical language modelling in the same direction as Resnik. Basili *et.al* continue their work [9] by developing a verb clustering algorithm which manages to identify semantically plausible classes of verb.

Church *et al.* [29] automatically parse a corpus into SVO triples. They then use this data to estimate mutual information collocations for SV, VS, SO, OS, VO, OV sub-pairs. The set of SV pairs identify, for example, typical verbs for a given subject: for the subject ⟨`boat`⟩, common collocations include ⟨`capsize`⟩, ⟨`sink`⟩, ⟨`cruise`⟩, ⟨`sail`⟩ and ⟨`tow`⟩.

Liddy and Paik [85] calculate the correlation between pairs of semantic tags supplied by the LDOCE dictionary using the correlation coefficient statistic. They include the information this provides into a hybrid system which contains heuristics for combining the multiple information sources.

We can consider the process of predicting particular semantic representations from a sentence to be similar to machine translation: in both cases, we are looking for the most likely alternative representation of a given input sentence. Speech recognition can also be viewed as the process of finding consistent mappings between an input representation of the sentence (acoustic) and a phonemic or orthographic output representation. This claim about underlying similarities is partly supported by the recent use researchers have made of parallel texts: in testing word sense disambiguation models, some researchers [51] have been using examples of text where the semantically ambiguous words are variously translated into two or more distinct foreign words, each of which broadly approximates one of the original word's senses. For particular sentences and particular semantically ambiguous words, the translated text affords an opportunity to evaluate the sense chosen by the disambiguation model. This testing method is appealing because it avoids reliance on expensive tagged corpora or lexicons which are sufficiently large. One of the difficulties with this method is an underlying assumption about the semantic importance of translation differences — just because there are three foreign words which approximately translate some word, it cannot be concluded that there are three distinct senses.

2.7 Machine Translation

In this section, we will give a brief overview of some recent developments in machine translation. Machine translation is, perhaps, the most difficult of all of the language processing specialisms. It requires the development of a robust model of the syntax of two languages (no practical model exists today even for a single language); useful machine translation systems will also require comprehensive semantic and possibly pragmatic and discourse-level models to avoid serious problems of disambiguation. Again, no detailed semantic description of a language has yet been developed. However, for those interested in constructing complete language processing systems, it offers a practical way of evaluating these systems whilst at the same time allowing them to carry out work on useful language applications.

Jones [70] describes a 'virtual translation' system as one which creates pairings between a source example and a target example, where examples are represented in a structure which includes morpho-

lexical, syntactic, semantic and pragmatic elements. Such a system makes an implicit assumption that underlying formalisms exist for each of these elements: for example, with lexical semantics, some language independent formalism needs to be used which can describe words from any language. Componential analysis may provide such a formalism, though there are many technical problems in their construction, and some criticisms of the theoretical assumptions which underpin them. Another system, which Jones uses, is functional grammar.

If fragments of text are considered as examples in Jones' system, then it can look for a match with some text fragment from another language by measuring their semantic similarity. Due to the emphasis on the meaning of text fragments (rather than, say, a detailed syntactic description), Jones argues that his system embodies a more *functional* view of language use than many of the language processing systems which we have presented so far. Jones uses versions of letter-based trigram language models, class-based models and a simple probabilistic word boundary detection system to describe the morphological and syntactic elements of text fragments, and implements a version of functional grammar manually. No details are given of the quality of the system.

Somers *et al.* [128] continue work on an example-based translation system. They need to use weighted n-grams to make better estimates of word alignment, because one of the languages they work with is German, which compounds some words more than English. They also use the automatic word classification system described by Schütze to tag words. The alignment of text fragments is now constrained by syntax rather than semantics as before. Recombination of translated fragments is constrained by the equivalent of a simplified probabilistic class-based n-gram model. Again, no detailed evaluation of the system is offered, though some translated fragments are presented.

Juola [72] describes another variation of example-based translation. He builds a system based on the 'marker hypothesis', which states that some (small) set of lexemes or morphemes exists in all natural languages and that these items are significant indicators of grammatical context. He embeds this hypothesis in a type of context-free grammar. He deals with lexical translation by maintaining multiple translation dictionaries, where each dictionary not only identifies the lexical mapping, but is associated with one particular syntactic context. Once source text fragments have been translated, the final target structure is recombined by a permutation of the source structure. Parameters are optimised by a simulated annealing algorithm. Results are reported from an experiment which was based on a corpus of 30 simplified bilingual sentences (generated by a grammar with a maximum of 10 rules and a lexicon of no more than 31 words) and tested on 47 sentences, 10 of which corresponded to novel syntactic structures. On the training data, Juola reports 36% correct, 21% minor errors and 44% gibberish. In [73], Juola subsumes his algorithm, which is based on the marker hypothesis, to a version of the categorial grammar [90] formalism. He describes an experiment which uses a bilingual corpus of sentences (no more than 7 words long) found in a child's reading book but reports disappionting results.

Gaussier *et al.* [53] use mutual information to discover bilingual word pairs from a bilinguial sentence-aligned corpus (from an idea proposed by Brown *et al.* [21]). Their system is less conceptually developed compared to Jones', though they report that 65% of word tokens are given 'good' assignments in the second language. Even a system which is so close to the data and with so bare a linguistic structure runs into sparse data problems.

Sutcliffe *et al.* [129] have also designed a lexical translation system. Their work could be considered as an implementation of one element (lexical semantic) of Jones' text fragments. Sutcliffe *et al.* map words into points in an n-dimensional common semantic space. The mapping is generated automatically from machine readable dictionaries to produce a kind of componential analysis of words. The definition of a word is parsed and its associated adjectives are added as semantic features (the

word 'furry', for example, becomes the semantic feature `+furry`). Two distinct sets of features can be extracted from two monolingual dictionaries and a mapping from the feature space of one language to the feature space of another is constructed manually. This type of approach has been described, for obvious reasons, as the *interlingua* approach. It contrasts with the directly *transfer*-based approach of, for example Gaussier *et al.*; however, the distinction may be transcended [39].

Also represented in the field of machine translation is the syntactic pattern recognition approach. Oncina *et al.* [104] present a learning algorithm which induces subsequential transducers — systems which translate sentences from one (formal) language to another, based on a finite state network which has transition arcs with associated input and output symbols. The authors offer results from an artificial translation task — visual descriptions of simple scenes, expressed via a context free language — of less that 0.1% error.

There are other recent grammar-based approaches to translation. Egedi *et al.* [39] describe a translation system based on a synchronous tree-adjoining grammar formalism. Whereas Resnik exploits lexicalised TAGs, Egedi *et al.* develop feature-based TAGs; here the features of the node where substitution takes place are unified with the features of the root node of the stustituting tree. Synchronous TAGs are simply pairs of grammars with a transfer rule-set mapping correspondences between nodes in the two grammars. If the two grammars capture the structure of two languages, then translation may be possible; if the two grammars capture the syntax and semantics of a single language, then a useful language generation/understanding system may be developed. The authors illustrate the benefits of their system by considering some typical linguistic difficulties: retalive clause, WH-questions, lexical selection (where they prefer selection-based constraints on unification rather than the interlingual approach seen in Sutcliffe *et al.*) and NP-recovery. Their synchronous TAG system is not stochastic.

3 Commonalities Between Language Processing Systems

Some of the statistical approaches described in this paper share common features with each other; they are also related to structure-discovering processes in other areas of cognitive science simulation, most apparently in some early research on visual processing. The underlying principles which unite this research come from information theory and statistics.

In the short period after Shannon's description of an information theoretic approach to language processing [124] and before Chomsky's influential criticism of finite models of language [28] psychologists investigated some of the powers and weaknesses of the distributional approach [95]. This work has continued, despite its lack of mainstream psycholinguistic appeal [96]. Contemporaneously, psychologists were also investigating the significance of an information-theoretic approach to the visual system; an early discussion of informational visual processing can be found in Attneave [3]. There has been no comparable rejection of information theory from cognitive models of visual processing.

In his article, Attneave spends some energy convincing the reader that the same principles of information theory are being used in language processing; he suggests that visual information is concentrated along contours, which fits well with the idea of entropy as a measure of uncertainty; Nicholis *et al.* [100] notice this phenomenon and offer it as evidence that a chaotic model might underly the perceptual and linguistic processes. The idea of identifying points of informational interest has some parallels with the recent idea of a distituent being associated with the point of minimum mutual information, described well in Brill *et al.* [17]; this idea is also supported by the pattern of prediction error rates in recurrent neural networks [40]. Attneave also links the psychological idea of *gestalt* with the information theoretic concept of high redundancy. In linguistics, this corresponds to utterances such as: ⟨`the cat sat the mat`⟩; the preposition is inferred by the hearer in order to construct a

meaningful sentence. Next, he states how these principles lead to a questioning of the connection between perception and inductive reasoning: that is, the boundary between perceptual information gathering and the central processing which the brain is supposed to perform. A convincing assault on the nature of this boundary has recently been presented by Dennett [36]. Attneave's main contribution in this article was to give some heuristic methods for reducing redundancy in an informational field. Many of these ideas can be subsumed by Algorithmic Complexity Theory [25, 127, 78]. Attneave also claims that something similar happens when good science is in operation [1].

Some connectionist researchers have described the functioning of their neural systems in terms of *information maximisation*, under certain constraints. Linsker [86] describes a multilayered neural net which uses Hebbian learning to self-organise feature-analysing cells. He states that the organising principle behind the changes in connection strength involves maximising the amount of information preserved in the signal as it moves through the layers. This is equivalent, given certain constraints, to maximising the statistical variance of each layer's output activity. Linsker not only shows that there is a relation between Hebbian and Hopfield networks and information theory, he also demonstrates the mathematical link between these connectionist learning rules and statistical variance. Cheng *et al.* [27] and Chater [26] offer some more detailed descriptions of the relationship between neural networks, probability theory and statistics.

The work of Finch *et al.* uses Spearman's rank correlation coefficient, ρ:

$$\rho = 1 - \frac{6 \sum_{j=1}^{k} D_j^2}{k^3 - k}$$

which can be shown [94] to be equivalent to

$$\rho = \frac{\sum_{j=1}^{k} (X_j - \overline{X})^2 (Y_j - \overline{Y})^2}{\sqrt{[\sum_{j=1}^{k} (X_j - \overline{X}) \sum_{j=1}^{k} (Y_j - \overline{Y})]}}$$

the correlation coefficient between two variables X and Y, which is just the covariance of their standard forms [22]. The closer this statistic is to 1, the stronger the correlation between the two sets of variables. Linsker has shown how covariance maximisation is related to information maximisation in a neural net; a related conclusion is that variable distributions which maximise the Spearman's rank correlation coefficient also maximise information.

These statistical and neural approaches suggest an underlying connection between both. Also, the 'replacement test' involves ideas which are obviously similar to the more general approach of estimating mutual information statistics, and to the 'variety index' of Faulk [44]; it is also similar to Brill *et al.*'s distributional analyses and Schütze's hybrid neural net and statistical clustering approach.

Similar links between neural network performance and information processing can be found in Plumbley [107] and Atick *et al.* [2], where the goal is described as minimisation of redundancy, corresponding to a minimisation of output channel capacity. A connection between information theory and artificial neuronal architectures is made by Gorin *et al* [58], who construct a network the weights of which are defined by mutual information.

Pereira *et al.* [105] cluster words using the Kullback-Leibler distance, or relative entropy. They use it to minimise the information loss in using a class distribution rather than the actual word distribution; mutual information is defined [34] as the relative entropy between the joint distribution,

[1] "The abstraction of simple homogeneities from a visual field does not appear to be different, in its formal aspects, from the induction of a highly general scientific law from a mass of experimental data"(Attneave [3], p187)

$P(X,Y)$ and the product distribution $P(X)P(Y)$; that is

$$M(X,Y) = D(P(X,Y) \mid\mid P(X)P(Y))$$

where $D(p_1 \mid\mid p_2)$ is the relative entropy between probability distributions p_1 and p_2; relative entropy as a measure of the distance between two distributions is closely related to covariance and the correlation coefficient.

In Kneser *et al.*'s system, we recall that they attempt to minimise the negative log-probability, LP of equation 5:

$$LP = -\sum_{g_1,g_2} N(g_1,g_2)\log N(g_1,g_2) + 2\sum_{g} N(g)\log N(g) - \sum_{w} N(w)\log N(w)$$

which is equivalent to

$$LP = -H(G_1,G_2) + 2H(G) - H(W) \tag{19}$$

Expanding $2H(G)$ with appropriate indices, we see that

$$LP = H(G_1) + H(G_2) - H(G_1,G_2) - H(W) \tag{20}$$

But $H(G_1)+H(G_2)-H(G_1,G_2)$ is just the mutual information between the two class sets, $M(G_1,G_2)$, so we conclude that finding an optimal negative log-likelihood is equivalent to maximising the average class mutual information (since $H(W)$ will not be influenced by word re-classifications).

Burger and Connolly [23] construct a Bayesian Network in the form of a tree and use sum squared error minimisation to calculate parameters for their system, which attempts to resolve anaphoric reference. This system is similar to Bahl *et al.*, who use entropy minimisation to build a decision tree for language modelling. Burger and Connolly derive their measure from the gradient descent of back-propagating neural networks, while Bahl *et al.* construct their language model equivalence classes by minimising the average entropy of leaf distributions — that is, they attempt to discover the maximally informative binary question at a tree-node. Their system can subsume word-based and class-based language models; it can also allow syntactic features to be included as context, just as probabilistic lexicalised tree-adjoining grammars can. Bahl *et al.*'s system can also include (in theory) semantic and pragmatic features of word context, as can Bod's data-oriented parsing scheme. Both these systems share some structural similarities with representations in example-based translation [97, 70].

Fisher and Riloff [50] use the t-statistic as a measure of co-occurrence likelihood between two items. It too is calculated from corpus frequency information and can indicate strong correlations between items. This statistic can measure collocational differences, whereas mutual information measures collocational similarities [29]. The likelihood ratio test allows measures of collocational similarity without assuming a normal distribution. Grefenstette [59] suggests that an adaptation of the Jaccard distance similarity measure leads to interesting language collocations — for example, his measure can be used to discover some interesting antonyms. The Jaccard measure is similar to Brill *et al.*'s distributional statistic; the measure is calculated as the fraction of shared features between two objects divided by the total number of their attributes. Ney, Essen and Kneser [99] include examples of word classification systems which, while not hierarchically clustered, use an optimisation technique based on decision-directed learning; their optimisation measure here is training set perplexity.

Statistical language processing techniques must deal with the phenomenon of sparse data; from n-gram language modelling to grammar induction, the poverty of available data has meant that the performance curves of many systems have flattened whenever they have been scaled up (to more complex grammars, or to larger vocabularies). This problem, however, is perennial in the field of artificial intelligence.

4 Evaluations of Language Processing Techniques

As an approximate rule of thumb, researchers who construct connectionist architectures tend to belong to the cognitive scientific school of corpus linguistics; those who use explicitly information theoretic measures tend to approach the subject from an engineering perspective. This division of approaches suggests some important differences when we examine how the various language processing systems are evaluated. Cognitive scientists look to psychology and linguistics for evaluation measures and engineers rely on measurable performance enhancement in particular applications (*e.g.* speech recognition and machine translation) for evaluation.

It should be clear from the section on automatic word classification that, though many of the methods may appear to use different approaches, there is a unifying concept into which most of the successful word classification systems can be transformed. This concept involves the quantification of the difference between two distributions — in this case, two distributions of word classes. Many successful word classification systems, to date, have worked by making operational definitions of the principles of structural linguistics. It remains to be seen, however, if these early successes can be improved upon sufficiently to make the structuralist approach any less unappealing to the mainstream of the linguistic community.

That these systems perform differently suggests that some measures are more appropriate than others; this highlights the need for discriminating system evaluation and also suggests a useful line of research in mathematical approaches to language: why are some models better performers? what does this suggest about building better mathematical theories of language? The power of the bigram statistic is anomalous and perhaps even surprising from a traditional theoretical linguistic point of view. Church *et al.* [29] have made a start on investigating the differences between various statistics which are commonly used in computational linguistics and lexicography. They explain the difference in lexical use between the t-test, from traditional statistics, and mutual information, from information theory. With mutual information, they claim, it is difficult to test for subtle dissimilarities between the use of two closely related words.

There are, in effect, no theoretical difficulties with engineering evaluations, only practical ones. This tends to make the reporting of results for particular systems less contarversial. Some practical disputes in evaluating systems include: assigning credit to almost-right translations and partial parses; evaluating the quality of the classification which results from an automatic word classification system; the proper treatment of unknown words in calculating the perplexity of unseen test sets.

On the other hand, not only must cognitive scientific researchers construct models which (measurably) work, but they must show that the systems work in a similar way to the human language processing system, or at least that the constructed model offers some other insight on human language processing. Unfortunately, there is still some dispute in psychology and linguistics over the plausibility of various models of language processing: the nature of the lexicon, the particular form of the (universal) grammar of languages, the relative autonomy of language processing sub-components and the nature of semantic and pragmatic processing, for example, all remain hotly debated issues.

Finally, we illustrate some difficulties faced by cognitive science researchers when they come to make conclusions about their systems; we take automatic word classification (and the corresponding cognitive scientific area — lexical classification in children) as an example. Children acquire language skills which include the ability to differentiate between types of word: ⟨`apple`⟩ is closer to ⟨`pear`⟩ than it is to ⟨`happy`⟩ or ⟨`sits`⟩, for example. Some word-classification systems can also (crudely) approximate this skill. Without any more information than this, we cannot bring the success of our word classification systems to bear on this cognitive scientific discussion. What extra information

could we add to make our computational model cognitively relevant?

It is sometimes implicitly assumed [106] that the ability to process natural language is so difficult that any extant demonstration of this ability must provide clues about *human* language processing. This stance is rarely supported by argument; it most certainly does not hold, for example with regard to the two ways that birds and planes fly. Redington *et al.* [113] eschew claims of direct cognitive relevance in their automatic word classification system, but they do hope that their system "..may have some bearing on the interpretation of behavioural data" (p1). Let us imagine, for the moment, a classification system which performs ideally: it fully classifies words, in all of their polysemic variety. However, children still might acquire their skill in an entirely different way. That both systems arrive at the same set of skills, of course, suggests similarities at a high level of generality, in the same way that cows and lions, at a high level of abstraction, gather food in similar ways, and that computer implementations of a context sensitive grammar and a context free grammar are similar (both are also finite-state grammars, by virtue of having been implemented on a computer) and that birds and planes fly in a similar way by virtue of using energy to counteract the force of gravity. Also, on some models of scientific activity [45], influence from the world of artificial intelligence can be just as inspiring as from any other area of cognitive scientists' lives.

If we assume a strong computational model of mind then a Turing machine exists which, given the same input as a child, can develop the same linguistic skills as a child. There may also be a Turing machine which can perform as well but with a smaller algorithmic complexity; indeed there may be just such a machine with the smallest algorithmic complexity possible, given the input data. It is unclear just how the complexity of the algorithm underlying the human achievement of this skill (innateness plus acquisition) compares to the minimum. Also, there may be other more powerful algorithms which can extract the same set of skills from a less rich input; and some weaker ones which require a more complex input to reach the same standard. In this area, the artificial intelligence approach can only deliver approximations to these algorithms; we need psycholinguistic experimentation to discover which are cognitively plausible.

References

[1] Hendrik James Antonisse. A grammar-based genetic algorithm. In *Foundations of Genetic Algorithms*, pages 193 – 204, 1991.

[2] Joseph J. Atick and A. Norman Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308 – 320, 1990.

[3] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183 – 193, 1954.

[4] Anthony F. Badalamenti, Robert J. Langs, and James Robinson. Lawful systems dynamics in how poets choose their words. *Behavioral Science*, 39, 1994.

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

[6] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 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 1983.

[7] Alwyn Barry. The emergence of high level structure in classifier systems — a proposal. In R. Cowie and M. Owens, editors, *Proceedings of the Sixth Irish Conference on Artificial Intelligence and Cognitive Science*, pages 185 – 196, September 1993.

[8] Roberto Basili, Teresa Pazienza, and Paolo Velardi. Combining NLP and statistical techniques for lexical acquisition. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

[9] Roberto Basili, Teresa Pazienza, and Paolo Velardi. What can be learned from raw texts? *Machine Translation*, 8:147 – 173, 1993.

[10] R. Beale and T. Jackson. *Neural Computing : An Introduction*. Adam Hilger, 1990.

[11] R. Beckwith, C. Fellbaum, D. Gross, and G. Miller. WordNet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*, chapter 9, pages 211 – 232. Lawrence Erlbaum Associates, 1991.

[12] Robert C. Berwick. Learning from positive-only examples — the subset principle and three case studies. In J. C. Carbonell R. S. Michalski and T. M. Mitchell, editors, *Machine Learning : An Artificial Intelligence Approach (Volume 2)*. Morgan Kaufmann Publishers, 1986.

[13] Rens Bod. A computational model of language performance. In *Proceedings* COLING-*92*, Nantes, 1992.

[14] Djamel Bouchaffra and Jacques Rouault. A nonstationary hidden markov model with a hard capture of observations : Application to the problem of morphological ambiguities. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

[15] Michael Brent. Minimal generative explanations : A middle ground between neurons and triggers. In *Proceedings of the Fifteenth Meeting of the Cognitive Science Society*, 1993.

[16] Eric Brill. Some advances in transformation-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI-94*, 1994.

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

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

[19] Rodney A. Brooks. Intelligence without reason. In *IJCAI-91*, 1991.

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

[21] P.F. Brown, J. Cocke, V. Della Pietra, S. Della Pietra, F. Jelinek, R.L. Mercer, and P.S. Roosin. A statistical approach to machine translation. *Computational Linguistics*, 16:79 – 85, 1990.

[22] K. A. Brownlee. *Statistical Theory and Methodology in Science and Engineering*. John Wiley and Sons inc., 1965.

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

[24] Glenn Carroll and Eugene Charniak. Learning probabilistic dependence grammars from labelled text. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

[25] G.J. Chaitin. Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos*, 4(1), February 1994.

[26] Nick Chater. Neural networks : The new statistical models of mind. In *Proceedings of the 1993 Neural Computation and Psychology Workshop*. UCL Press, London, 1994.

[27] Bing Cheng and D. M. Titterington. Neural networks : A review from a statistical perspective. *Statistical Science*, 9(1):2 – 54, 1994.

[28] Noam Chomsky. *Syntactic Structures*. The Hague: Mouton, 1957.

[29] K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*, chapter 6, pages 115 – 164. Lawrence Erlbaum Associates, 1991.

[30] Kenneth W. Church and William A. Gale. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech and Language*, 5:19 – 54, 1991.

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

[32] Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on applied Natural Language processing*, 1988.

[33] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, pages 76 – 82, 1989.

[34] Thomas M. Cover and Joy A. Thomas. *Elements of Information theory.* John Wiley and Sons, 1991.

[35] Ferdinand de Saussure. *Course in General Linguistics.* Duckworth, 1983.

[36] Daniel Dennett. *Consciousness Explained.* London: Allen Lane, 1991.

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

[38] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61 – 74, 1993.

[39] D. Egedi, M. Palmer, H.S. Park, and A.K. Joshi. Korean to English translation using synchronous TAGs. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 48 – 55, Columbia, Maryland, October 1994.

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

[41] Jeffrey L. Elman. Incremental learning, or the importance of starting small. Technical Report 9101, Center for Research in Language, U.C.S.D., 1991.

[42] David Elworthy. Does baum-welch re-estimation help taggers? In *Proceedings of the fourteenth ACL Conference on Applied Natural Language Processing, ANLP-94*, pages 53 – 58, October 1994.

[43] R. Fano. *Transmission of Information.* M.I.T. Press, 1961.

[44] R. D. Faulk and F. Goertzel Gustavson. Segmenting discrete data representing continuous speech input. *I.B.M. Systems Journal*, 29(2), 1990.

[45] Paul Feyerabend. *Against Method: Outline of an Anarchist Theory of Knowledge.* NLB, London, 1975.

[46] Steven Finch and Nick Chater. 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, 1992.

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

[48] Steven Paul Finch. *Finding Structure in Language.* PhD thesis, Centre for Cognitive Science, University of Edinburgh, 1993.

[49] J. R. Firth. A synopsis of linguistic theory 1930 – 1955. In F. Palmer, editor, *Selected Papers of J. R. Firth.* Longman, 1968.

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

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

[52] Michael Gasser. Learning syllable representations : A connectionist approach. In *The Cognitive Science of Natural Language Processing*, 1992.

[53] Eric Gaussier and Jean-Marc Lange. Towards bilingual terminology. In *Proceedings of the 19th International Conference of the Association for Literary and Linguistic Computing and the 12th International Conference on Computers and the Humanities*, pages 121 – 124, 1992.

[54] M. Gold. Language identification in the limit. *Information and Control*, 10:447 – 474, 1967.

[55] D.E. Goldberg. *Genetic Algorithms in Search Optimization and Machine Learning.* Addison Wesley, 1989.

[56] Rafael C. Gonzalez and Michael G. Thomason. *Syntactic Pattern Recognition.* Addison Wesley, 1978.

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

[58] A.L. Gorin, S.E. Levinson, A.N. Gertner, and E. Goldman. Adaptive acquisition of language. *Computer Speech and Language*, 5:101 – 132, 1991.

[59] Gregory Grefenstette. Finding semantic similarity in raw text : The Deese antonyms. In *Probabilistic Approaches to Natural Language.* American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

[60] Zellig S. Harris. *Structural Linguistics.* Phoenix Books, 1951.

[61] J. H. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, 1975.

[62] J. H. Holland. Escaping brittleness : The possibilities of general purpose learning algorithms applied to parallel rule-bases systems. In J. G. Carbonell R. S. Michalski and T. M. Mitchell, editors, *Machine Learning II.* Morgan Kaufmann, 1986.

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

[64] John Hughes and Eric Atwell. A methodical approach to word class formation using automatic evaluation. Presented at The Society of Artificial Intelligence and Simulation of Behaviour, April 1994.

[65] Jonathan J. Hull. Combining syntactic knowledge and visual text recognition : A hidden markov model for part of speech tagging in a word recognition algorithm. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

[66] Frederick Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the I.E.E.E.*, 64(4), April 1976.

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

[68] Frederick Jelinek. Self-organized language modelling for speech recognition. In Waibel and Lee, editors, *Readings in Speech recognition*. Morgan Kaufmann. San Mateo, California, 1990.

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

[70] Daniel Jones. Virtual machine translation. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

[71] M. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 531 – 546. Lawrence Erlbaum Associates, 1986.

[72] Patrick Juola. A psycholinguistic approach to corpus-based machine translation. In A.I.C. Monaghan, editor, *Third Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, 1994.

[73] Patrick Juola. Corpus-based acquisition of grammars and transfer functions for machine translation. Technical report CU-CS-756-95, Department of Computer Science, University of Colorado at Boulder, 1995.

[74] Patrick Juola, Chris Hall, and Adam Boggs. Corpus-based morphological segmentation by entropy changes. In A.I.C. Monaghan, editor, *Third Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, 1994.

[75] Slava M. Katz. 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 1987.

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

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

[78] A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems in Information Transmission*, 1:4 – 7, 1964.

[79] J. R. Koza. Hierarchical genetic algorithms that operate on populations of computer programs. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 768 – 780, 1989.

[80] Hideki Kozima. Text segmentation based on similarity between words. In *Proceedings of the Association for Computational Linguistics*, 1993.

[81] Alexander Krotov, Robert Gaizauskas, and Yorick Wilks. Acquiring a stochastic context-free grammar from the penn treebank. In A. I. C. Monaghan, editor, *Third Conference on the Cognitive Science of Natural Language Processing*, Dublin City University, 1994.

[82] Ronald Kuhn and Renato De Mori. 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 1990.

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

[84] Mark K. Liberman. The trend towards statistical models in natural language processing. In E. Klein and F. Veltman, editors, *Natural Language and Speech*. Springer verlag, Berlin, 1991.

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

[86] Ralph Linsker. Self-organization in a perceptual network. *I.E.E.E. Computer*, 21(3):105 – 117, 1988.

[87] David M. Magerman. *Natural Language Parsing as Statistical Pattern Recognition*. PhD thesis, Stanford University Computer Science Department, February 1994.

[88] John Makhoul, Fred Jelinek, Larry Rabiner, Clifford Weinstein, and Victor Zue. Spoken language systems. *Annual Review of Computer Science*, 4:481 – 501, 1990.

[89] Bernard Manderick. The genetic algorithm. In *Background and Experiments in Machine Learning of Natural Language*, 1992.

[90] M. McGee-Wood. *Categorial Grammars*. Routledge, London, 1993.

[91] John McMahon. *Statistical Language processing Based on Self-Organising Word Classification*. PhD thesis, Department of Computer Science, Queen's University of Belfast, 1994.

[92] John McMahon and F. J. Smith. Improving statistical language model performance with automatically generated word hierarchies. in Preparation.

[93] John McMahon and F. J. Smith. Structural tags, annealing and automatic word classification. in Press : Artificial Intelligence and the Simulation of Behaviour Quarterly, 1995.

[94] Ray Meddis. *Statistics Using Ranks — A Unified Approach*. Basil Blackwell, 1984.

[95] George A. Miller. *Language and Communication*. New York: McGraw-Hill, 1951.

[96] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1 – 28, 1991.

[97] Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Barnerji, editors, *Artificial and Human Intelligence*, pages 173 – 180. North-Holland, 1984.

[98] Sven Naumann and Jürgen Schrepp. Inductive learning of reversible grammars. In Walter Daelemans and David Powers, editors, *Background and Experiments in Machine Learning of Natural Language*, pages 237–243. Institute for Language Technology and AI, 1992.

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

[100] John S. Nicholis and Anastassis A. Katsikas. Chaotic dynamics of linguistic-like processes at the syntactical and semantic levels : In pursuit of a multifractal attractor. In *Patterns, Information and Chaos in Neuronal Systems*. World Scientific, 1993.

[101] Peter O'Boyle. *A Study of an N-gram Language Model for Speech Recognition*. PhD thesis, Department of Computer Science, Queen's University, Belfast, 1993.

[102] Peter O'Boyle, Marie Owens, and F.J. Smith. A study of a statistical model of natural language. *The Irish Journal of Psychology*, 14(3):382 – 396, 1993.

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

[104] J. Oncina, A. Castellanos, E. Vidal, and V. Jimémez. Corpus-based machine translation through subsequential transducers. In A.I.C. Monaghan, editor, *Third Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, 1994.

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

[106] Steven Pinker. *The Language Instinct — The New Science of Language and Mind*. Allen Lane, Penguin Press, 1994.

[107] M. D. Plumbley. Information theory and neural network learning algorithms. In Gerry Orchard, editor, *Neural Computing – Research and Applications*, pages 145 – 155. Institute of Physics Publishing, 1993. Proceedings of the Second Irish Neural Networks Conference.

[108] J. B. Pollack. Recursive distributed representations. *Artificial Intelligence*, 46:77 – 105, 1990.

[109] David Powers and Walter Daelemans. SHOE : The extraction of hierarchical structure for machine learning of natural language (project summary). In *Background and Experiments in Machine Learning of Natural Language*, pages 125–161, 1992.

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

[111] Allan Ramsay. Linguistics : The cognitive science of natural language. In *Third Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, July 1994.

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

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

[114] Ronan Reilly. A connectionist technique for on-line parsing. In *The Cognitive Science of Natural Language Processing*, 1992.

[115] Ronan Reilly. An exploration of clause boundary effects in simple recurrent network representations. In *The Second Irish Neural Netowrks Conference*, 1992.

[116] Philip S. Resnik. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of* COLING-*92*, Nantes, 1992.

[117] Philip S. Resnik. *Selection and Information : A Class-Based Approach to Lexical Relationships*. PhD thesis, Computer and Information Science, University of Pennsylvania, December 1993. Institute for Research in Cognitive Science Report I.R.C.S.-93-42.

[118] Jan Robin Rohlicek, Yen-Lu Chow, and Salim Roucos. Sytatistical language modelling using a small corpus from an application domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 267 – 270, 1988.

[119] Geoffrey Sampson. 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*, pages 219 – 226. Rodopi,Amsterdam, 1987.

[120] J. C. Scholtes. Resolving linguistic ambiguities with a neural data oriented parsing (dop) system. In *Background and Experiments in Machine Learning of Natural Language*, pages 279–282, 1992.

[121] Hinrich Schütze. Context space. In *Probabilistic Approaches to Natural Language*. American Association for Artificial Intelligence, AAAI Press, 1992. Technical report FS-92-05.

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

[123] Hinrich Schütze and Jan Pedersen. A vector model for syntagmatic and paradagmatic relatedness. To Appear in 'Making Sense of Words : Proceedings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research.

[124] C.E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 1951.

[125] D. Solomon. Learning a grammar. Technical Report UMCS-AI-91-12-1, University of Manchester Department of Computer Science, 1991.

[126] D. Solomon and M. McGee-Wood. Unified lexicon and grammar. In Russell J. Collingham, editor, *Workshop on the Unified Lexicon*, December 1993.

[127] R. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1 – 22 and 224 – 254, 1964.

[128] H. Somers, I. McLean, and D. Jones. Experiments in multilingual example-based generation. In A.I.C. Monaghan, editor, *Third Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, 1994.

[129] Richard F.E. Sutcliffe, Annette McElligott, and G. O'Néill. Irish-English lexical translation using distributed semantic representations. In R. Cowie and M. Owens, editors, *Artificial Intelligence and Cognitive Science*, 1993.

[130] Michael K. Tanenhaus. Psycholinguistics : An overview. In Frederick J. Newmeyer, editor, *Linguistics : The Cambridge Survey*, volume III, chapter 1, pages 1–37. Cambridge University Press, 1988.

[131] M. Mitchell Waldrop. *Complexity : The Emerging Science at the Edge of Order and Chaos*. Viking, 1993.

[132] J. Gerard Wolff. *Towards a Theory of Cognition and Computing*. Ellis Horwood, 1991.

[133] J. Gerard Wolff. Language learning, cognition and computing : A summary. In *Background and Experiments in Machine Learning of Natural Language*, 1992.

[134] P. J. Wyard and C. Nightingale. Grammar recognition by a single layer higher order neural net. *B.T. Technological Journal*, 10(3), 1992.

[135] Uri Zernik. Introduction. In Uri Zernik, editor, *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*, chapter 1, pages 1 – 26. Lawrence Erlbaum Associates, 1991.

[136] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.