Introduction to the Special Issue on Computational Linguistics Using Large Corpora

Kenneth W. Church*
AT&T Bell Laboratories

Robert L. Mercer†
IBM T.J. Watson Research Center

The 1990s have witnessed a resurgence of interest in 1950s-style empirical and statistical methods of language analysis. Empiricism was at its peak in the 1950s, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory). At that time, it was common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Firth, a leading figure in British linguistics during the 1950s, summarized the approach with the memorable line: "You shall know a word by the company it keeps" (Firth 1957). Regrettably, interest in empiricism faded in the late 1950s and early 1960s with a number of significant events including Chomsky's criticism of n-grams in *Syntactic Structures* (Chomsky 1957) and Minsky and Papert's criticism of neural networks in *Perceptrons* (Minsky and Papert 1969).

Perhaps the most immediate reason for this empirical renaissance is the availability of massive quantities of data: more text is available than ever before. Just ten years ago, the one-million word Brown Corpus (Francis and Kučera, 1982) was considered large, but even then, there were much larger corpora such as the Birmingham Corpus (Sinclair et al. 1987; Sinclair 1987). Today, many locations have samples of text running into the hundreds of millions or even billions of words. Collections of this magnitude are becoming widely available, thanks to data collection efforts such as the Association for Computational Linguistics' Data Collection Initiative (ACL/DCI), the European Corpus Initiative (ECI), ICAME, the British National Corpus (BNC), the Linguistic Data Consortium (LDC), the Consortium for Lexical Research (CLR), Electronic Dictionary Research (EDR), and standardization efforts such as the Text Encoding Initiative (TEI).[1] The data-intensive approach to language, which is becoming known as *Text Analysis*, takes a pragmatic approach that is well suited to meet the recent emphasis on numerical evaluations and concrete deliverables. Text Analysis focuses on broad (though possibly superficial) coverage of unrestricted text, rather than deep analysis of (artificially) restricted domains.

* AT&T Bell Laboratories, Office 2B-421, 600 Mountain Ave., Murray Hill, NJ 07974.
† IBM T.J. Watson Research Center, P.O. Box 704, J2-H24, Yorktown Heights, NY 10598.

[1] For more information on the ACL/DCI, contact Felicia Hurewitz, ACL/DCI, Room 619, Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305, USA, (tel) 215-898-0083, (fax) 215-573-2091, (e-mail) fel@unagi.cis.upenn.edu. For more information on the LDC, contact Elizabeth Hodas, Linguistic Data Consortium, Room 441, Williams Hall, University of Pennsylvania, Philadelphia, PA, 19104-6305, USA, (tel) 215-898-0464, (fax) 215-573-2175, (e-mail) ehodas@unagi.cis.upenn.edu. Send e-mail to smbowie@vax.oxford.ac.uk for information on the BNC, to lexical@nmsu.edu for information on the CLR, and to eucorp@cogsci.edinburgh.ac.uk for information on the ECI. Information on the London-Lund Corpus and other corpora available through ICAME can be found in the ICAME Journal, edited by Stig Johansson, Department of English, University of Oslo, Norway.

The case for the resurgence of empiricism in computational linguistics is nicely summarized in Susan Warwick-Armstrong's call-for-papers for this special issue:

> The increasing availability of machine-readable corpora has suggested new methods for studies in a variety of areas such as lexical knowledge acquisition, grammar construction, and machine translation. Though common in the speech community, the use of statistical and probabilistic methods to discover and organize data is relatively new to the field at large. The various initiatives currently under way to locate and collect machine-readable corpora have recognized the potential of using this data and are working toward making these materials available to the research community. Given the growing interest in corpus studies, it seems timely to devote an issue of CL to this topic.

In Section 1, we review the experience of the speech recognition community. Stochastic methods based on Shannon's noisy channel model have become the methods of choice within the speech community. Knowledge-based approaches were tried during the first DARPA speech recognition project in the early 1970s, but have largely been abandoned in favor of stochastic approaches that have become the main focus of DARPA's more recent efforts.

In Section 2, we discuss how this experience is influencing the language community. Many of the most successful speech techniques are achieving major improvements in performance in the language area. In particular, probabilistic taggers based on Shannon's noisy channel model are becoming the method of choice because they correctly tag 95% of the words in a new text, a major improvement over earlier technologies that ignored lexical probabilities and other preferences that can be estimated statistically from corpus evidence.

In Section 3, we discuss a number of frequency-based preferences such as *collocations* and *word associations*. Although often ignored in the computational linguistics literature because they are difficult to capture with traditional parsing technology, they can' easily overwhelm syntactic factors (as any psycholinguist knows). Four articles in this special issue take a first step toward preference-based parsing, an empirical alternative to the rational tradition of principle-based parsing, ATNs, unification, etc.

In Section 4, we discuss entropy and evaluation issues, which have become relatively important in recent years.

In Section 5, we discuss the application of noisy channel models to bilingual applications such as machine translation and bilingual lexicography.

In Section 6, we discuss the use of empirical methods in monolingual lexicography, contrasting the exploratory data analysis (EDA) view of statistics with other perspectives such as hypothesis testing and supervised/unsupervised learning/training. There are five articles in this special issue on computational lexicography, using both the exploratory and the self-organizing approaches to statistics.

1. The Influence from the Speech Community

1.1 Consensus in Speech Community: Stochastic Methods Are Outperforming Knowledge-Based Methods

Over the past 20 years, the speech community has reached a consensus in favor of empirical methods. As observed by Waibel and Lee in the introduction to their collection of reprints on speech recognition (Waibel and Lee 1990):

> Chapter 5 describes the *knowledge-based approach,* proposed in the 1970s and early 1980s. The pure knowledge-based approach emulates human speech knowledge using expert systems. Rule-based systems have had only limited success... Chapter 6 describes the *stochastic approach...* Most successful large-scale systems today use a stochastic approach. (Waibel and Lee 1990; p. 4)

A number of data collection efforts have helped to bring about this change in the speech community, especially the Texas Instruments' Digit Corpus (Leonard 1984), TIMIT and the DARPA Resource Management (RM) Database (Price et al. 1988). According to the Linguistic Data Consortium (LDC), the RM database was used by every paper that reported speech recognition results in the *1988 Proceedings of IEEE ICASSP,* the major technical society meeting where speech recognition results are reported. This is especially significant given that abstracts for this meeting were due just a few months after the release of the corpus, attesting to the speech recognition community's hunger for standard corpora for development and evaluation.

Back in the 1970s, the more data-intensive methods were probably beyond the means of many researchers, especially those working in universities. Perhaps some of these researchers turned to the knowledge-based approach because they couldn't afford the alternative. It is an interesting fact that most of the authors of the knowledge-based papers in Chapter 5 of Waibel and Lee (1990) have a university affiliation whereas most of the authors of the data-intensive papers in Chapter 6 have an industrial affiliation. Fortunately, as a result of improvements in computer technology and the increasing availability of data due to numerous data collection efforts, the data-intensive methods are no longer restricted to those working in affluent industrial laboratories.

1.2 The Anti-Empiricist Period in Speech Research

At the time, of course, the knowledge-based approach was not advocated on economic grounds. Rather, the knowledge-based approach was advocated as necessary in order to deal with the lack of allophonic invariance. The mapping between phonemes and their allophonic realizations is highly variable and ambiguous. The phoneme /t/, for example, may be realized as a released stop in "Tom," as a flap in "butter," or as a glottal stop in "bottle." Two different phonemes may lead to the same allophonic variant in some contexts. For example, "writer" and "rider" are nearly identical in many dialects of American English. Residual differences, such as the length of the preconsonantal vowel, are easily overwhelmed by the context in which the word appears. Thus, if one says "Joe is a rider of novels," listeners hear "Joe is a writer of novels," while if one says "Joe is a writer of horses," listeners hear "Joe is a rider of horses." Listeners usually have little problem with the wild variability and ambiguity of speech because they know what the speaker is likely to say.

> In most systems for sentence recognition, such modifications must be viewed as a kind of 'noise' that makes it more difficult to hypothesize lexical candidates given an input phonetic transcription. To see that this must be the case, we note that each phonological rule [in the utterance: "Did you hit it to Tom?"] results in irreversible ambiguity—the phonological rule does not have a unique inverse that could be used to recover the underlying phonemic representation for a lexical

> item. For example, ... [t]he tongue flap ... could have come from a /t/ or a /d/. (Klatt 1980; pp. 548–549)

The first DARPA Speech Understanding project emphasized the use of high-level constraints (e.g., syntax, semantics, and pragmatics) as a tool to disambiguate the allophonic information in the speech signal by understanding the message. At BBN, researchers called their system HWIM for (*Hear What I Mean*). They hoped to use NLP techniques such as ATNs (Woods 1970) to understand the sentences that they were trying to recognize even though the output of their front end was highly variable and ambiguous.

The emphasis today on empirical methods in the speech recognition community is a reaction to the failure of knowledge-based approaches of the 1970s. It has become popular once again to focus on high-level natural language constraints in order to reduce the search space. But this time, n-gram methods have become the methods of choice because they seem to work better than the alternatives, at least when the search space is measured in terms of entropy. Ideally, we might hope that someday parsers might reduce entropy beyond that of n-grams, but right now, parsers seem to be more useful for other tasks such as understanding who did what to whom, and less useful for predicting what the speaker is likely to say.

1.3 The Raleigh System: A Foundation for the Revival of Empiricism

In the midst of all of this excitement over high-level, knowledged-based NLP techniques, IBM formed a new speech group around the nucleus of an existing group that was moved from Raleigh, North Carolina, to Yorktown Heights early in 1972. The Raleigh group brought to Yorktown a working speech recognition system, that had been designed in accordance with prevailing anti-empiricist attitudes of the time, though it would soon serve as a foundation for the revival of empiricism in the speech and language communities.

The front end of the Raleigh system converted the speech signal (20,000 samples per second) first into a sequence of 80 filter bank outputs (100 80-dimensional vectors per second), and then into a sequence of phoneme-like labels (100 labels per second), using an elaborate set of hand-tuned rules that would soon be replaced with an automatically trained procedure. The back end converted these labels into a sequence of words using an artificial finite-state grammar that was so small that the finite-state machine could be written down on a single piece of paper. Today, many systems attempt to model unrestricted language using methods that will be discussed in Section 3, but at the time, it was standard practice to work with artificial grammars of this kind.

When it worked perfectly, the front end produced a transcription of the speech signal such as might be produced by a human phonetician listening carefully to the original speech. Unfortunately, it almost never worked perfectly, even on so small a stretch as a single word. Rapid phones, such as flaps, were often missed; long phones, such as liquids and stressed vowels, were often broken into several separate segments; and very often phones were simply mislabeled. The back end was designed to overcome these problems by navigating through the finite-state network, applying a complicated set of hand-tuned penalties and bonuses to the various paths in order to favor those paths where the low-level acoustics matched the high-level grammatical constraints. This system of hand-tuned penalties and bonuses correctly recognized 35% of the sentences (and 77% of the words) in the test set. At the time, this level of performance was actually quite impressive, but these days, one would expect much more, now that most systems use parameters trained on real data, rather than a complicated set of hand-tuned penalties and bonuses.

1.4 The Noisy Channel Model

Although the penalties and bonuses were sometimes thought of as probabilities, the early Raleigh system lacked a complete and unified probabilistic framework. In a radical departure from the prevailing attitudes of the time, the Yorktown group turned to Shannon's theory of communication in the presence of noise and recast the speech recognition problem in terms of transmission through a noisy channel. Shannon's theory of communication (Shannon 1948), also known as *Information Theory*, was originally developed at AT&T Bell Laboratories to model communication along a noisy channel such as a telephone line. See Fano (1961) for a well-known secondary source on the subject, or Cover and Thomas (1991) or Bell, Cleary, and Witten (1990) for more recent treatments.

The noisy channel paradigm can be applied to other recognition applications such as optical character recognition (OCR) and spelling correction. Imagine a noisy channel, such as a speech recognition machine that almost hears, an optical character recognition (OCR) machine that almost reads, or a typist who almost types. A sequence of good text (I) goes into the channel, and a sequence of corrupted text (O) comes out the other end.

$$I \rightarrow \textit{Noisy Channel} \rightarrow O$$

How can an automatic procedure recover the good input text, I, from the corrupted output, O? In principle, one can recover the most likely input, $\hat{I}$, by hypothesizing all possible input texts, I, and selecting the input text with the highest score, $Pr(I \mid O)$. Symbolically,

$$\hat{I} = \underset{I}{\text{argmax}} Pr(I \mid O) = \underset{I}{\text{argmax}} Pr(I)\ Pr(O \mid I)$$

where ARGMAX finds the argument with the maximum score.

The prior probability, $Pr(I)$, is the probability that I will be presented at the input to the channel. In speech recognition, it is the probability that the talker utters I; in spelling correction (Damerau 1964; Kukich 1992), it is the probability that the typist intends to type I. In practice, the prior probability is unavailable, and consequently, we have to make do with a model of the prior probability, such as the trigram model. The parameters of the language model are usually estimated by computing various statistics over a large sample of text.

The channel probability, $Pr(O \mid I)$, is the probability that O will appear at the output of the channel when I is presented at the input; it is large if I is similar, in some appropriate sense, to O, and small, otherwise. The channel probability depends on the application. In speech recognition, for example, the output for the word "writer" may look similar to the word "rider"; in character recognition, this will not be the case. Other examples are shown in Table 1.

1.5 Training Is Better than Guessing

Rather than rely on guesses for the values of the bonuses and penalties as the Raleigh group had done, the Yorktown group used three levels of hidden Markov models (HMMs) to compute the conditional probabilities necessary for the noisy channel. A Markov model is a finite state machine with probabilities governing transitions between states and controlling the emission of output symbols. If the sequence of state transitions cannot be determined when the sequence of outputs is known, the Markov model is said to be *hidden*. In practice, the Forward–Backward algorithm is often used to estimate the values of the transition and emission parameters on the basis of corpus evidence. See Furui (1989; Appendix D.3, pp. 343–347) for a brief description of the

Table 1
Examples of channel confusions in different applications.

Application	Input	Output
Speech	writer	rider
Recognition	here	hear
Optical	all	all (*A-one-L*)
Character	of	o{
Recognition	form	farm
Spelling	government	goverment
Correction	occurred	occured
	commercial	commerical

Table 2
Performance after training (Bahl et al. 1975)

Training Set Size	Test Sentences Correctly Decoded	Decoding Problems
0	2/10	8/10
200	77/100	3/100
400	80/100	2/100
600	85/100	1/100
800	82/100	3/100
1070	83/100	3/100

Forward–Backward algorithm, and (Rabiner 1989) for a longer tutorial on HMMs. The general procedure, of which the Forward–Backward algorithm is a special case, was first published and shown to converge by Baum (1972).

The first level of the Raleigh system converted spelling to phonemic base forms, rather like a dictionary; the second level dealt with the problems of allophonic variation mentioned above; the third level modeled the front end. At first, the values of the parameters in these HMMs were carefully constructed by hand, but eventually they would all be replaced with estimates obtained by training on real data using statistical estimation procedures such as the Forward–Backward algorithm.

The advantages of training are apparent in Table 2. Note the astounding improvement in performance. Despite a few decoding problems, which indicate limitations in the heuristic search procedure employed by the recognizer, sentence accuracy had improved from 35% to 82–83%.

Moreover, training turned out to be important for speeding up the search. The first row shows the results for the initial estimates, which were very carefully prepared by two members of the group over several weeks. Despite all of the careful hand work, the search was so slow that only 10 of the 100 test sentences could be recognized. The initial estimates were unusable without at least some training. These days, most researchers find that they do not need to be nearly so careful in obtaining initial estimates.

Emboldened by this success, the group began to explore other areas where training might be helpful. They began by throwing out the phonological rules. Thus, they accepted only a single pronunciation for each word. *B-u-t-t-e-r* had to be pronounced *butter* and *s-o-m-e-t-h-i-n-g* had to be pronounced *something,* and that was that. Any change in these pronunciations was treated as a mislabeling from the front end. After training, this simplified system correctly decoded 75% of 100 test sentences, which was very encouraging.

Finally, they removed the dictionary lookup HMM, taking for the pronunciation of each word its spelling. Thus, a word like *t-h-r-o-u-g-h* was assumed to have a pronunciation like *tuh huh ruh oh uu guh huh.* After training, the system learned that with words like *l-a-t-e* the front end often missed the *e*. Similarly, it learned that *g*'s and *h*'s were often silent. This crippled system was still able to recognize 43% of 100 test sentences correctly as compared with 35% for the original Raleigh system.

These results firmly established the importance of a coherent, probabilistic approach to speech recognition and the importance of data for estimating the parameters of a probabilistic model. One by one, pieces of the system that had been assiduously assembled by speech experts yielded to probabilistic modeling. Even the elaborate set of hand-tuned rules for segmenting the frequency bank outputs into phoneme-sized segments would be replaced with training (Bakis 1976; Bahl et al. 1978). By the summer of 1977, performance had reached 95% correct by sentence and 99.4% correct by word, a considerable improvement over the same system with hand-tuned segmentation rules (73% by sentence and 95% by word).

Progress in speech recognition at Yorktown and almost everywhere else as well has continued along the lines drawn in these early experiments. As computers increased in power, ever greater tracts of the heuristic wasteland opened up for colonization by probabilistic models. As greater quantities of recorded data became available, these areas were tamed by automatic training techniques. Today, as indicated in the introduction of Waibel and Lee (1990), almost every aspect of most speech recognition systems is dominated by probabilistic models with parameters determined from data.

2. Part-of-Speech Tagging

Many of the very same methods are being applied to problems in natural language processing by many of the very same researchers. As a result, the empirical approach has been adopted by almost all contemporary part-of-speech programs: Bahl and Mercer (1976), Leech, Garside, and Atwell (1983), Jelinek (1985), Derouault and Merialdo (1986), Garside, Leech, and Sampson (1987), Church (1988), DeRose (1988), Hindle (1989), Kupiec (1989, 1992), Ayuso et al. (1990), deMarcken (1990), Karlsson (1990), Boggess, Agarwal, and Davis (1991), Merialdo (1991), and Voutilainen, Heikkila, and Anttila (1992). These programs input a sequence of words, e.g., *The chair will table the motion,* and output a sequence of part-of-speech tags, e.g., *art noun modal verb art noun.* Most of these programs correctly tag at least 95% of the words, with practically no restrictions on the input text, and with very modest space and time requirements. Perhaps the most important indication of success is that many of these statistical tagging programs are now being used on large volumes of data (hundreds of millions of words of text) in a number of different application areas including speech synthesis (Sproat, personal communication; Liberman and Church 1991), speech recognition (Jelinek 1985; Jelinek, Mercer, and Roukos 1991), information retrieval (Salton, Zhao, and Buckley 1990; Croft, Turtle, and Lewis 1991), sense disambiguation (Hearst 1991), and computational lexicography (Klavans and Tzoukermann 1990a, 1990b; Church and Gale 1991). Apparently, these programs must be addressing some important needs of the research community or else they wouldn't be as widely cited as they are. Many of the papers in this special issue refer to these taggers.

As in speech recognition, data collection efforts have played a pivotal role in advancing data-intensive approaches to part-of-speech tagging. The Brown Corpus (Francis and Kučera 1982) and similar efforts within the ICAME community, have created invaluable opportunities. The Penn Treebank (see the paper by Marcus and Santorini 1983) is currently being distributed by the ACL/DCI. The European Corpus Initiative (ECI) plans to distribute similar material in a variety of languages. Even greater resources are expected from the Linguistic Data Consortium (LDC). And the Consortium for Lexical Research (CLR) is helping to make dictionaries more accessible to the research community. For information on contacting these organizations, see footnote 1.

2.1 Recasting Part-of-Speech Tagging as a Noisy Channel Problem

Many of the tagging programs mentioned above are based on Shannon's Noisy Channel Model. Imagine that a sequence of parts of speech, P, is presented at the input to the channel and for some crazy reason, it appears at the output of the channel in a corrupted form as a sequence of words, W. Our job is to determine P given W.

$$P \rightarrow \textit{Noisy Channel} \rightarrow W$$

By analogy with the noisy channel formulation of the speech recognition problem, the most probable part-of-speech sequence, $\hat{P}$, is given by:

$$\hat{P} = \underset{P}{\operatorname{argmax}} Pr(P)\ Pr(W \mid P)$$

In theory, with the proper choice for the probability distributions $Pr(P)$ and $Pr(W \mid P)$, this algorithm will perform as well as, or better than, any possible alternative that one could imagine. Unfortunately, the probability distributions $Pr(P)$ and $Pr(W \mid P)$ are enormously complex: $Pr(W \mid P)$ is a table giving for every pair W and P of the same length a number between 0 and 1 that is the probability that a sequence of words chosen at random from English text and found to have the part-of-speech sequence P will turn out to be the word sequence W. Changing even a single word or part-of-speech in a long sequence may change this number by many orders of magnitude. However, experience has shown that surprisingly high tagging accuracy can be achieved in practice using very simple approximations to $Pr(P)$ and $Pr(W \mid P)$. In particular, it is possible to replace $Pr(P)$ by a trigram approximation:

$$Pr(P_1, P_2, \ldots, P_N) \approx \prod_{i=i}^{N} Pr(P_i \mid P_{i-2}P_{i-1})$$

and to replace $Pr(W \mid P)$ by an approximation in which each word depends only on its own part of speech:

$$Pr(W_1, W_2, \ldots, W_N \mid P_1, P_2, \ldots, P_N) \approx \prod_{i=i}^{N} Pr(W_i \mid P_i)$$

In these equations, P_i is the i^{th} part of speech in the sequence P, and W_i is the i^{th} word in W. The parameters of this model, the lexical probabilities, $Pr(W_i \mid P_i)$, and the contextual probabilities, $Pr(P_i \mid P_{i-2}P_{i-1})$, are generally estimated by computing various statistics over large bodies of text. One can view the first set of parameters as a dictionary and the second set of parameters as a grammar.

Table 3
Lexical ambiguity is hard (if we ignore preferences).

Word	Part-of-Speech Tags	
	More Likely	Less Likely
I	pronoun	noun (letter of alphabet)
see	verb	noun (*the Holy See*)
a	article	noun (letter of alphabet)
bird	noun	verb

2.2 Why Traditional Methods Have Failed

Traditional methods have tended to ignore lexical preferences, which are the single-most important source of constraint for part-of-speech tagging, and are sufficient by themselves to resolve 90% of the tags. Consider the trivial sentence, "I see a bird," where every word is almost unambiguous. In the Brown Corpus (Francis and Kučera 1982), the word "I" appears as a pronoun in 5,131 times out of 5,132 ($\approx$ 100%), "see" appears as a verb in 771 times out of 772 ($\approx$ 100%), "a" appears as an article in 22,938 times out of 22,944 ($\approx$ 100%) and "bird" appears as a noun in 25 times out of 25 ($\approx$ 100%). However, in addition to the desired tag, many dictionaries such as Webster's Ninth New Collegiate Dictionary (Mish 1983) also list a number of extremely rare alternatives, as illustrated in Table 3. These alternatives can usually be eliminated on the basis of the statistical preferences, but traditional parsers don't, and consequently run into serious difficulties. Attempts to eliminate unwanted tags on syntactic grounds have not been very successful. For example, *I/noun see/noun a/noun bird/noun,* cannot be ruled out as syntactically ill-formed, because the parser must accept sequences of four nouns in other situations: *city school committee meeting.* Apparently, syntactic rules are not nearly as effective as lexical preferences, at least for this application.

The tradition of ignoring preferences dates back to Chomsky's introduction of the competence approximation (Chomsky 1957, pp. 15–17). Recall that Chomsky was concerned that approximations, such as Shannon's n-gram approximation, which was very much in vogue at the time, were inappropriate for his needs, and therefore, he introduced an alternative with complementary strengths and weaknesses. The competence approximation is more appropriate for modeling long-distance dependences such as agreement constraints and wh-movement, but at the cost of missing certain crucial local constraints, especially the kinds of preferences that are extremely important for part-of-speech tagging.

2.3 Using Statistics to Fit Probabilistic Models to Data

Probabilistic models provide a theoretical abstraction of language, very much like Chomsky's competence model. They are designed to capture the more important aspects of language and ignore the less important aspects, where what counts as important depends on the application. Statistics are often used to estimate the values of the parameters in these probabilistic models. Thus, for example, we might estimate the probability distribution for the word *Kennedy* in the Brown Corpus by modeling the distribution with a *binomial,* and then use the frequency of *Kennedy* in the Brown Corpus (140) to fit the model to the data.

The classic example of a binomial process is coin tossing. Suppose that the coin comes up heads with probability p. Then the probability that it will come up heads exactly m times in n tosses is

$$\binom{n}{m} p^m (1-p)^{n-m}.$$

Here

$$\binom{n}{m},$$

which is called the *binomial coefficient,* is the number of ways the m positions can be chosen from the n coin tosses. It is equal to

$$\frac{n!}{m!(n-m)!},$$

where $n!$ (n factorial) is equal to $1 \times 2 \times \cdots \times n$. For example, tossing a fair coin three times ($n = 3$, $p = 1/2$) will result in 0, 1, 2, and 3 heads with probability 1/8, 3/8, 3/8, and 1/8, respectively. This set of probabilities is called the *binomial distribution* for n and p. The expected value of the binomial distribution is np and the variance is $\sigma^2 = np(1-p)$. Thus, tossing a fair coin three times will produce an average of 3/2 heads with a variance of 3/4.

How can the binomial be used to model the distribution of *Kennedy*? Let p be the probability that a word chosen at random in English text is *Kennedy*. We can think of a series of words in English text as analogous to tosses of a coin that comes up heads with probability p: the coin is heads if the word is *Kennedy*, and is tails otherwise. Of course, we don't really know the value of p, but in a sample of n words, we should expect to find about np occurrences of *Kennedy*. There are 140 occurrences of *Kennedy* in the Brown Corpus, for which n is approximately 1,000,000. Therefore, we can argue that $1,000,000p$ must be about 140 and we can make an estimate, $\hat{p}$, of p equal to $140/1,000,000$. If we really believe that words in English text come up like heads when we flip a biased coin, then $\hat{p}$ is the value of p that makes the Brown Corpus as probable as possible. Therefore, this method of estimating parameters is called *maximum likelihood estimation* (MLE). For simple models, MLE is very easy to implement and produces reasonable estimates in many cases. More elaborate methods such as the *Good-Turing Method* (Good 1953) or *Deleted Estimation* (Jelinek and Mercer 1980, 1985) should be used when the frequencies are small (e.g., less than 10).

It is often convenient to use these statistical estimates as if they are the same as the true probabilities, but this practice can lead to trouble, especially when the data don't fit the model very well. In fact, content words don't fit a binomial very well, because content words tend to appear in "bursts." That is, content words are like buses in New York City; they are social animals and like to travel in packs. In particular, if the word *Kennedy* appears once in a unit of text (e.g., a paragraph, a discourse, or a genre), then it is much more likely than chance to appear a second time in the same unit of text. Function words also deviate from the binomial, though for different reasons (e.g., stylistic factors mentioned in Biber's paper).

These bursts might serve a useful purpose. People seem to be able to use these bursts to speed up reaction times in various tasks. Psycholinguists use the term *priming* to refer to this effect. Bursts might also be useful in a number of practical applications such as Information Retrieval (IR) (Salton 1989; Frakes and Baeza-Yates 1992). There have been a number of attempts over the years to model these bursts. The negative binomial distribution, for example, was explored in considerable detail in the classic study of the authorship of the Federalist Papers, Mosteller and Wallace (1964), a must-read for anyone interested in statistical analyses of large corpora.

We can show that the distribution of *Kennedy* is very bursty in the Brown Corpus by dividing the corpus into k segments and showing that the probability varies radically from one segment to another. For example, if we divide the Brown Corpus into 10 segments of 100,000 words each, we find that the frequency of *Kennedy* is: 58, 57, 2, 12, 6, 1, 4, 0, 0, 0. The variance of these 10 numbers is 539. Under the binomial assumption, we obtain a very different estimate of the variance. In a sample of $n = 100,000$ words, with $\hat{p} = 140$ per million, we would expect a variance of $n\hat{p}(1-\hat{p}) \approx 14$. (The variance of the binomial is approximately the same as the expected value when p is small.) The large discrepancy between the empirically derived estimate of the variance (539) and the one based on the binomial assumption (14) indicates that the binomial assumption does not fit the data very well.

3. Preference-Based Parsers

When the data don't fit the model very well, we may wish to look for alternative models. Four articles in this special issue propose empirical alternatives to traditional parsing methods based on the competence model. As we have seen, the competence model doesn't fit the part-of-speech application very well because of the model's failure to capture certain lexical preferences. The model also runs into trouble in a number of other NLP applications. Consider, for example, the problem of deciding between the words *form* and *farm* in the OCR application (mentioned in Section 1.4) when they appear in the context:

$$pure \begin{pmatrix} farm \\ form \end{pmatrix} of$$

Most people would have little difficulty deciding that *form* was the intended word. Neither does an OCR system that employs a trigram language model, because preferences, such as *collocations,* fall naturally within the scope of the n-gram approximation. Traditional NLP techniques, on the other hand, fail here because the competence approximation does not capture the crucial collocational constraints.

Lexicographers use the terms *collocation, co-occurrence,* and *lexis* to describe various constraints on pairs of words. The words *strong* and *powerful* are perhaps the canonical example. Halliday (1966; p. 150) noted that although *strong* and *powerful* have similar syntax and semantics, there are contexts where one is much more appropriate than the other (e.g., *strong tea* vs. *powerful computers*).

Psycholinguists have a similar concept, which they call *word associations.* Two frequently cited examples of highly associated words are: *bread/butter* and *doctor/nurse.* See Palermo and Jenkins (1964) for tables of associations, measured for 200 words, factored by grade level and sex. In general, subjects respond more quickly to a word such as *butter* when it follows a highly associated word such as *bread.*

> Some results and implications are summarized from reaction-time experiments in which subjects either (a) classified successive strings of letters as words and nonwords, or (b) pronounced the strings. Both types of response to words (e.g., BUTTER) were consistently faster when preceded by associated words (e.g., BREAD) rather than unassociated words (e.g, NURSE). (Meyer, Schvaneveldt, and Ruddy 1975; p. 98)

Table 4
The trigram approximation in action (Jelinek 1985).

Word	Rank	More likely alternatives
We	9	The This One Two A Three Please In
need	7	are will the would also do
to	1	
resolve	85	have know do ...
all	9	the this these problems ...
of	2	the
the	1	
important	657	document question first ...
issues	14	thing point to ...

These constraints are rarely discussed in computational linguistics because they are not captured very well with traditional NLP techniques, especially those based on the competence approximation. Of course, it isn't hard to build computational models that capture at least some of these preferences. Even the trigram model, despite all of its obvious shortcomings, does better than many traditional methods in this regard. The power of the trigram approximation is illustrated in Table 4 for the sentence fragment, *We need to resolve all of the important issues...*, selected from a 90 million–word corpus of IBM office correspondences. Each row shows the correct word, the rank of the correct word as predicted by the trigram model, and then the list of words judged by the trigram model to be more probable than the correct word. Thus, *We* is the 9th most probable word to begin a sentence. At this point in the sentence, in the absence of any other context, the trigram model is as good as any model we could have. Following *We* at the beginning of the sentence, *need* is the 7th most probable word, ranking behind *are, will, the, would, also,* and *do*. Here, again, the trigram model still accounts for all of the context there is and so should be doing as well as any model can. Following *We need, to* is the most probable word. Although by now, the trigram model has lost track of the complete context (it no longer realizes that we are at the beginning of a sentence), it is still doing very well.

Table 4 shows that the trigram model captures a number of important frequency-based constraints that would be missed by most traditional parsers. For example, the trigram model captures the fact that *issues* is particularly predictable in the collocation: *important issues*. In general, high-frequency function words like *to* and *the*, which are acoustically short, are more predictable than content words like *resolve* and *important*, which are longer. This is convenient for speech recognition because it means that the language model provides more powerful constraints just when the acoustic model is having the toughest time. One suspects that this is not an accident, but rather a natural result of the evolution of speech to fill the human needs for reliable communication in the presence of noise.

. The ideal NLP model would combine the strengths of both the competence approximation and the n-gram approximation. One possible solution might be the *Inside–Outside algorithm* (Baker 1979; Lari and Young 1991), a generalization of the Forward–Backward algorithm that estimates the parameters of a hidden stochastic context-free grammar, rather than a hidden Markov model. Four alternatives are proposed in these special issues: (1) Brent (1993), (2) Briscoe and Carroll (this issue), (3) Hindle and Rooth (this issue), and (4) Weischedel et al. (1993). Briscoe and Carroll's contribution is very much in the spirit of the Inside–Outside algorithm, whereas Hindle and Rooth's contribution, for example, takes an approach that is much closer to the concerns of lexicography, and makes use of preferences involving words, rather than preferences that ignore words and focus exclusively on syntactic structures. Hindle and Rooth show how co-occurrence statistics can be used to improve the performance of the parser on sentences such as:

$$\text{She} \begin{pmatrix} \text{wanted} \\ \text{placed} \\ \text{put} \end{pmatrix} \text{the dress on the rack.}$$

where lexical preferences are crucial to resolving the ambiguity of prepositional phrase attachment (Ford, Bresnan, and Kaplan 1982). Hindle and Rooth show that a parser can enforce these preferences by comparing the statistical association of the verb-preposition (*want...on*) with the association of the object-preposition (*dress...on*), when attaching the prepositional phrase. This work is just a first step toward preference-based parsing, an empirically motivated alternative to traditional rational approaches such as ATNs, unification parsers, and principle-based parsers.

4. Entropy

How do we decide if one language model is better than another? In the 1940s, Shannon defined *entropy*, a measure of the information content of a probabilistic source, and used it to quantify such concepts as noise, redundancy, the capacity of a communication channel (e.g., a telephone), and the efficiency of a code. The standard unit of entropy is the *bit* or *bi*nary digi*t*. See Bell, Cleary, and Witten (1990) for a more discussion on entropy; Section 2.2.5 shows how to compute the entropy of a model, and Section 4 discusses how Shannon and others have estimated the entropy of English.

From the point of view of speech recognition or OCR, we would like to be able to characterize the size of the search space, the number of binary questions that the recognizer will have to answer on average in order to decode a message. *Cross entropy* is a useful yardstick for measuring the ability of a language model to predict a source of data. If the language model is very good at predicting the future output of the source, then the cross entropy will be small. No matter how good the language model is, though, the cross entropy cannot be reduced below a lower bound, known as the *entropy* of the source, the cross entropy of the source with itself.

One can also think of the cross entropy between a language model and a probabilistic source as the number of bits that will be needed on average to encode a symbol from the source when it is assumed, albeit mistakenly, that the language model is a perfect probabilistic characterization of the source. Thus, there is a close connection between a language model and a coding scheme. Table 5 below lists a number of coding schemes along with estimates of their cross entropies with English text.

The standard ASCII code requires 8 bits per character. It would be a perfect code if the source produced each of the $2^8 = 256$ symbols equally often and independently of context. However, English is not like this. For an English source, it is possible to reduce the average length of the code by assigning shorter codes to more frequent symbols (e.g., *e*, *n*, *s*) and longer codes to less frequent symbols (e.g., *j*, *q*, *z*), using a coding scheme such as a Huffman code (Bell, Cleary, and Witten 1990; Section 5.1.2). Other codes, such as Lempel–Ziv (Welch 1984; Bell, Cleary, and Witten, Chapters 8–9) and n-gram models on words, achieve even better compression by taking advantage of context, though none of these codes seem to perform as well as people do in predicting the next letter (Shannon 1951).

Table 5
Cross entropy of various language models.

Model	Bits / Character
ASCII	8
Huffman code each char	5
Lempel-Ziv (Unix™ *compress*)	4.43
Unigram (Huffman code each word)	2.1 (Brown, personal communication)
Trigram	1.76 (Brown et al. 1992)
Human Performance	1.25 (Shannon 1951)

The cross entropy, H, of a code and a source is given by:

$$H(source, code) = -\sum_{s}\sum_{h} Pr(s, h \mid source) \log_2 Pr(s \mid h, code)$$

where $Pr(s, h \mid source)$ is the joint probability of a symbol s following a history h given the source. $Pr(s \mid h, code)$ is the conditional probability of s given the history (context) h and the code. In the special case of ASCII, where $Pr(s \mid h, ASCII) = 1/256$, we can actually carry out the indicated sum, and find, not surprisingly, that ASCII requires 8 bits per character:

$$H(source, ASCII) = -\sum_{s=1}^{256} \frac{1}{256} \log_2 \frac{1}{256} = 8$$

In more difficult cases, cross entropy is estimated by a sampling procedure. Two independent samples of the source are collected: S_1 and S_2. The first sample, S_1, is used to fit the values of the parameters of the code, and second sample, S_2, is used to test the fit. For example, to determine the value of 5 bits per character for the Huffman code in Table 5, we counted the number of times that each of the 256 ASCII characters appeared in S_1, a sample of N_1 characters selected from the *Wall Street Journal* text distributed by the ACL/DCI. These counts were used to determine $Pr(s \mid h, code)$ (or rather $Pr(s \mid code)$, since the Huffman code doesn't depend on h). Then we collected a second sample, S_2, of N_2 characters, and tested the fit with the formula:

$$H(source, code) \approx -\frac{1}{N_2} \sum_{i=1}^{N_2} \log_2 Pr(S_2[i] \mid code)$$

where $S_2[i]$ is the i^{th} character in the second sample. It is important in this procedure to use two different samples of text. If we were to use the same sample for both testing and training, we would obtain an overly optimistic estimate of how well the code performs.

The other codes in Table 5 make better use of context (h), and therefore, they achieve better compression. For example, Huffman coding on words (a unigram model) is more than twice as compact as Huffman coding on characters (2.1 vs. 5 bits/char.). The unigram model is also more than twice as good as Lempel–Ziv (2.1 vs. 4.43 bits/char.), demonstrating that *compress*, a popular Unix™ tool for compressing files, could

Table 6
Summary of two approaches to NLP.

	Rationalism	Empiricism
Well-known Advocates:	Chomsky, Minsky	Shannon, Skinner, Firth, Harris
Model:	Competence Model	Noisy Channel Model
Contexts of Interest:	Phrase Structure	N-grams
Goals:	All and Only Explanatory Theoretical	Minimize Prediction Error (Entropy) Descriptive Applied
Linguistic Generalizations:	Agreement and Wh-movement	Collocations and Word Associations
Parsing Strategies:	Principle-Based CKY (Chart), ATNs, Unification	Preference-Based Forward–Backward, Inside–Outside
Applications:	Understanding Who did what to whom	Recognition Noisy Channel Applications

be improved by a factor of two (when the files are in English). The trigram model, the method of choice in speech recognition, achieves 1.76 bits per character, outperforming the practical alternatives in Table 5, but falling half a bit shy of Shannon's estimate of human performance.[2]

Someday parsers might help squeeze out some of this remaining half bit between the trigram model and Shannon's bound, but thus far, parsing has had little impact. Lari and Young (1991; p. 255), for example, conducted a number of experiments with stochastic context-free grammars (SCFGs), and concluded that "[t]he experiments on word recognition showed that although SCFGs are effective, their complex training routine prohibits them from directly replacing the simpler HMM-based recognizers." They then proceeded to argue, quite sensibly, that parsers are probably more appropriate for tasks where phrase structure is more directly relevant than in word recognition. In general, phrase structure is probably more important for understanding who did what to whom, than recognizing what was said.[3] Some tasks are probably more appropriate for Chomsky's rational approach to language and other tasks are probably more appropriate for Shannon's empirical approach to language. Table 6 summarizes some of the differences between the two approaches.

5. Machine Translation and Bilingual Lexicography

Is machine translation (MT) more suitable for rationalism or empiricism? Both approaches have been investigated. Weaver (1949) was the first to propose an information theoretic approach to MT. The empirical approach was also practiced at Georgetown during the 1950s and 1960s (Henisz-Dostert, Ross Macdonald, and Zarechnak 1979) in a system that eventually became known as SYSTRAN. Recently, most work in MT

[2] In fact, the trigram model might be even better than suggested in Table 5, since the estimate for the trigram model in Brown et al. (1992) is computed over a 256-character alphabet, whereas the estimate for human performance in Shannan (1951) is computed over a 27-character alphabet.

[3] Lari and Young actually looked at another task involving phonotactic structure where there is also good reason to believe that SCFGs might be able to capture crucial linguistic constraints that might be missed by simpler HMMs.

has tended to favor rationalism, though there are some important exceptions, such as example-based MT (Sato and Nagao 1990). The issue remains as controversial as ever, as evidenced by the lively debate on rationalism versus empiricism at TMI-92, a recent conference on MT.[4]

The paper by Brown et al. (1990) revives Weaver's information theoretic approach to MT. It requires a bit more squeezing and twisting to fit machine translation into the noisy channel mold: to translate, for example, from French to English, one imagines that the native speaker of French has thought up what he or she wants to say in English and then translates mentally into French before actually saying it. The task of the translation system is to recover the original English, E, from the observed French, F. While this may seem a bit far-fetched, it differs little in principle from using English as an interlingua or as a meaning representation language.

$$E \longrightarrow \textit{Noisy Channel} \longrightarrow F$$

As before, one minimizes one's chance of error by choosing E according to the formula:

$$\hat{E} \;=\; \underset{E}{\operatorname{argmax}} Pr(E)\; Pr(F \mid E)$$

As before, the parameters of the model are estimated by computing various statistics over large samples of text. The prior probability, $Pr(E)$, is estimated in exactly the same way as discussed above for the speech recognition application. The parameters of the channel model, $Pr(F \mid E)$, are estimated from a parallel text that has been aligned by an automatic procedure that figures out which parts of the source text correspond to which parts of the target text. See Brown et al. (1993) for more details on the estimation of the parameters.

The information theoretic approach to MT may fail for reasons advanced by Chomsky and others in the 1950s. But regardless of its ultimate success or failure, there is a growing community of researchers in corpus-based linguistics who believe that it will produce a number of lexical resources that may be of great value. In particular, there has been quite a bit of discussion of bilingual concordances recently (e.g., Klavans and Tzoukermann 1990a, 1990b; Church and Gale 1991), including the 1990 and 1991 lexicography conferences sponsored by Oxford University Press and Waterloo University. A bilingual concordance is like a monolingual concordance except that each line in the concordance is followed by a line of text in a second language. There are also some hopes that the approach might produce tools that could be useful for human translators (Isabelle 1992).

There are three papers in these special issues on aligning bilingual texts such as the Canadian Hansards (parliamentary debates) that are available in both English and French: Brown et al. (1993), Gale and Church (this issue), and Kay and Rösenschein (this issue). Warwick-Armstrong and Russell have also been interested in the alignment problem (Warwick-Armstrong and Russell 1990). Except for Brown et al., this work is focused on the less controversial applications in lexicography and human translation, rather than MT.

[4] Requests for a tape of the debate should be sent to the attention of Pierre Isabelle, CCRIT, TMI-92, 1575 boul. Chomedey, Laval (Quebec), H7V 2X2, Canada. Copies of the TMI proceedings can be obtained by writing to CCRIT or sending e-mail to tmi@ccrit.doc.ca.

6. Monolingual Lexicography, Machine-Readable Dictionaries (MRDs), and Computational Lexicons

There has been a long tradition of empiricist approaches in lexicography, both bilingual and monolingual, dating back to Johnson and Murray. As corpus data and machine-readable dictionaries (MRDs) become more and more available, it is becoming easier to compile lexicons for computers and dictionaries for people. This is a particularly exciting area in computational linguistics as evidenced by the large number of contributions in these special issues: Biber (1993), Brent (1993), Hindle and Rooth (this issue), Pustejovsky et al. (1993), and Smadja (this issue). Starting with the COBUILD dictionary (Sinclair et al. 1987), it is now becoming more and more common to find lexicographers working directly with corpus data. Sinclair makes an excellent case for the use of corpus evidence in the preface to the COBUILD dictionary:

> For the first time, a dictionary has been compiled by the thorough examination of a representative group of English texts, spoken and written, running to many millions of words. This means that in addition to all the tools of the conventional dictionary makers—wide reading and experience of English, other dictionaries and of course eyes and ears—this dictionary is based on hard, measurable evidence. (Sinclair et al. 1987; p. xv)

The experience of writing the COBUILD dictionary is documented in Sinclair (1987), a collection of articles from the COBUILD project; see Boguraev (1990) for a strong positive review of this collection. At the time, the corpus-based approach to lexicography was considered pioneering, even somewhat controversial; today, quite a number of the major lexicography houses are collecting large amounts of corpus data.

The traditional alternative to corpora are citation indexes, boxes of interesting citations collected on index cards by large numbers of human readers. Unfortunately, citation indexes tend to be a bit like butterfly collections, full of rare and unusual specimens, but severely lacking in ordinary, garden-variety moths. Murray, the editor of the *Oxford English Dictionary*, complained:

> The editor or his assistants have to search for precious hours for examples of common words, which readers passed by... Thus, of *Abusion* we found in the slips about 50 instances; of *Abuse* not five. (James Augustus Henry Murray, Presidential Address, *Philological Society Transactions 1877–9*, pp. 571–2, quoted by Murray 1977, p. 178)

He then went on to say, "There was not a single quotation for *imaginable*, a word used by Chaucer, Sir Thomas More, and Milton." From a statistical point of view, citation indexes have serious sampling problems; they tend to produce a sample that is heavily skewed away from the "central and typical" facts of the language that every speaker is expected to know. Large corpus studies, such as the COBUILD dictionary, offer the hope that it might be possible to base a dictionary on a large and representative sample of the language as it is actually used.

6.1 Should a Corpus Be Balanced?

Ideally, we would like to use a large and representative sample of general language, but if we have to choose between large and representative, which is more important? There was a debate on a similar question between Prof. John Sinclair and Sir Randolf

Table 7
Coverage of *imaginable* in various corpora.

Size (in millions)	Corpus	raw freq	freq/million
1	Brown Corpus	0	0
1	Bible	0	0
2	Shakespeare	0	0
7	WSJ	41	5.9
10	Groliers	5	0.5
18	Hansard	15	0.8
29	DOE	5	0.2
46	AP 1988	36	0.8
50	AP 1989	39	0.8
56	AP 1990	21	0.4
47	AP 1991	19	0.4

Quirk at the 1991 lexicography conference sponsored by Oxford University Press and Waterloo University, where the house voted, perhaps surprisingly, that a corpus does not need to be balanced. Although the house was probably predisposed to side with Quirk's position, Sinclair was able to point out a number of serious problems with the balancing position. It may not be possible to properly balance a corpus. And moreover, if we insist on throwing out idiosyncratic data, we may find it very difficult to collect any data at all, since all corpora have their quirks.

In some sense, the question comes down to a tradeoff between quality and quantity. American industrial laboratories (e.g., IBM, AT&T) tend to favor quantity, whereas the BNC, NERC, and many dictionary publishers, especially in Europe, tend to favor quality. The paper by Biber (1993) argues for quality, suggesting that we ought to use the same kinds of sampling methods that statisticians use when studying the economy or predicting the results of an election. Poor sampling methods, inappropriate assumptions, and other statistical errors can produce misleading results: "There are lies, damn lies, and statistics."

Unfortunately, sampling methods can be expensive; it is not clear whether we can justify the expense for the kinds of applications that we have in mind. Table 7 might lend some support for the quantity position for Murray's example of *imaginable.* Note that there is plenty of evidence in the larger corpora, but not in the smaller ones. Thus, it would appear that "more data are better data," at least for the purpose of finding exemplars of words like *imaginable.*

Similar comments hold for collocation studies, as illustrated in Table 8, which shows mutual information values (Fano 1961; p. 28)

$$I(x;y) = \log_2 \frac{Pr(x,y)}{Pr(x)\, Pr(y)}$$

for several collocations in a number of different corpora. Mutual information compares the probability of observing word x and word y *together* (the joint probability) to the probability of observing x and y *independently* (chance). Most of the mutual information values in Table 8 are much larger than zero, indicating, as we would hope, that the collocations appear much more often in these corpora than one would expect by chance. The probabilities, $Pr(x)$ and $Pr(y)$, are estimated by counting the number of observations of x and y in a corpus, $f(x)$ and $f(y)$, respectively, and normalizing by N,

Table 8
Collocations in different corpora.

Size (in millions)	Corpus	*strong* *support*	*strong* *economy*	*strong* *winds*	*strong* *man*	*strong* *enough*
1	Brown Corpus	5.1 (1)	—	8.3 (1)	—	7.3
1	Bible	—	—	—	3.4 (7)	—
2	Shakespeare	—	—	—	—	6.5 (4)
7	WSJ	5.5	4.9	6.5 (7)	4.7	6.5
10	Groliers	5.8	3.8 (3)	8.3	4.2 (3)	8.3
18	Hansard	6.2	6.4	—	—	7.0
29	DOE	4.5	4.3 (4)	7.7	—	7.4
46	AP 1988	6.3	6.3	8.5	4.0	7.0
50	AP 1989	6.3	4.7	8.4	1.8 (7)	7.3
56	AP 1990	6.5	3.7	8.3	2.4 (9)	7.5
47	AP 1991	7.0	3.4	8.7	2.0 (6)	7.3

the size of the corpus. The joint probability, $Pr(x,y)$, is estimated by counting the number of times that x is immediately followed by y in the corpus, $f(x,y)$, and normalizing by N. Unfortunately, mutual information values become unstable if the counts are too small. For this reason, small counts (less than 10) are shown in parentheses. A dash is used when there is no evidence for the collocation.

Like Table 7, Table 8 also shows that "more data are better data." That is, there is plenty of evidence in the larger corpora, but not in the smaller ones. "Only a large corpus of natural language enables us to identify recurring patterns in the language and to observe collocational and lexical restrictions accurately. . . ." (Hanks 1990; p. 36)

However, in order to make use of this evidence we have to find ways to compensate for the obvious problems of working with unbalanced data. For example, in the Canadian Hansards, there are a number of unwanted phrases such as: "House of Commons," "free trade agreement," "honour and duty to present," and "Hear! Hear!" Fortunately, though, it is extremely unlikely that these unwanted phrases will appear much more often than chance across a range of other corpora such as Department of Energy (DOE) abstracts or the Associated Press (AP) news. If such a phrase were to appear relatively often across a range of such diverse corpora, then it is probably worthy of further investigation. Thus, it is not required that the corpora be balanced, but rather that their quirks be uncorrelated across a range of different corpora. This is a much weaker and more realistic requirement than the more standard (and more idealistic) practice of balancing and purging quirks.

6.2 Lexicography and Exploratory Data Analysis (EDA)

Statistics can be used for many different purposes. Traditionally, statistics such as Student's t-tests were developed to *test* a particular hypothesis. For example, suppose that we were concerned that *strong enough* shouldn't be considered a collocation. A t-test could be used to compare the hypothesis that *strong enough* appears too often to be a fluke against the *null hypothesis* that the observations can be attributed to chance. The t-score compares the two hypotheses, by taking the difference of the means of the two probability distributions, and normalizing appropriately by the variances, so that the result can be interpreted as a number of standard deviations. Theoretically, if the t-score is larger than 1.65 standard deviations, then we ought to believe that the co-occurrences are *significant* and we can reject the null hypothesis with 95% confidence, though in practice we might look for a t-score of 2 or more standard deviations, since t-scores are often inflated (due to certain violations of the assumptions behind the model). See Dunning (this issue) for a critique of the assumption that the probabilities are normally distributed, and an alternative parameterization of the probability distributions.

$$t = \frac{mean(Pr(strong, enough)) - mean(Pr(strong))\ mean(Pr(enough))}{\sqrt{\sigma^2(Pr(strong, enough)) + \sigma^2(Pr(strong)Pr(enough))}}$$

In the Brown Corpus, it happens that $f(strong, enough) = 11$, $f(strong) = 194$, $f(enough) = 426$, and $N = 1,181,041$. Using these values, we estimate $t \approx 3.3$, which is larger than 1.65, and therefore we can confidently reject the null hypothesis, and conclude that the co-occurrence is *significantly* larger than chance. The estimation uses the approximation, $\sigma^2(Pr(strong, enough)) \approx f(strong, enough)/N^2$, which can be justified under appropriate binomial assumptions. It is also assumed that $\sigma^2(Pr(strong)\ Pr(enough))$ is very small and can be omitted.

$$t \approx \frac{\frac{f(strong,enough)}{N} - \frac{f(strong)}{N}\frac{f(enough)}{N}}{\sqrt{\frac{f(strong,enough)}{N^2}}} \approx 3.3$$

Although statistics are often used to test a particular hypothesis as we have just seen, statistics can also be used to *explore* the space of possible hypotheses, or to *discover* new hypotheses (supervised/unsupervised learning/training). See Tukey (1977) and Mosteller and Tukey (1977) for two textbooks on Exploratory Data Analysis (EDA), and Jelinek (1985) for a very nice review paper on self-organizing statistics. Both the exploratory and self-organizing views are represented in these special issues. Pustejovsky et al. (1993) use an EDA approach to investigate certain questions in lexical semantics. Brent (1993), in contrast, adopts a self-organizing approach to identify subcategorization features.

Table 9 shows how the t-score can be used in an exploratory mode to extract large numbers of words from the Associated Press (AP) news that co-occur more often with *strong* than with *powerful*, and vice versa. It is an interesting question whether collocations are simply idiosyncratic as Halliday and many others have generally assumed (see Smadja [this issue]), or whether there might be some general principles that could account for many of the cases. After looking at Table 9, Hanks, a lexicographer and one of the authors of Church et al. (1991), hypothesized that strong is an intrinsic quality whereas *powerful* is an extrinsic one. Thus, for example, any worthwhile politician or cause can expect *strong supporters*, who are enthusiastic, convinced, vociferous, etc., but far more valuable are *powerful supporters*, who will bring others with them. They are also, according to the AP news, much rarer—or at any rate, much less often mentioned. This is a fascinating hypothesis that deserves further investigation.

Summary statistics such as mutual information and t-scores may have an important role to play in helping lexicographers to discover significant patterns of collocations, though the position remains somewhat controversial. Some lexicographers prefer mutual information, some prefer t-scores, and some are unconvinced that either of them is any good. Church et al. (1991) argued that different statistics have different strengths and weaknesses, and that it requires human judgment and exploration to decide which statistic is best for a particular problem. Others, such as Jelinek (1985), would prefer a self-organizing approach, where there is no need for human judgment.

Table 9
An example of the *t*-score (Church et al. 1991).

	Strong w				Powerful w		
t	strong w	powerful w	w	t	strong w	powerful w	w
12.42	161	0	showing	−7.44	1	56	than
11.94	175	2	support	−5.60	1	32	figure
10.08	550	68	,	−5.37	3	31	minority
9.97	106	0	defense	−5.23	1	28	of
9.76	102	0	economy	−4.91	0	24	post
9.50	97	0	demand	−4.63	5	25	new
9.40	95	0	gains	−4.35	27	36	military
9.18	91	0	growth	−3.89	0	15	figures
8.84	137	5	winds	−3.59	6	17	presidency
8.02	83	1	opposition	−3.57	27	29	political
7.78	67	0	sales	−3.33	0	11	computers

7. Conclusion

The flourishing renaissance of empiricism in computational linguistics grew out of the experience of the speech recognition community during the 1970s and 1980s. Many of the same statistical techniques (e.g., Shannon's Noisy Channel Model, n-gram models, hidden Markov models (HMMs), entropy (H), mutual information (I), Student's *t*-score) have appeared in one form or another, often first in speech, and then soon thereafter in language. Many of the same researchers have applied these methods to a variety of application areas ranging from language modeling for noisy channel applications (e.g., speech recognition, optical character recognition [OCR], and spelling correction [Damerau 1964; Kukich 1992]), to part-of-speech tagging, parsing, translation, lexicography, text compression (Bell, Cleary, and Witten 1990) and information retrieval (IR) (Salton 1989; Frakes and Baeza-Yates 1992).

Empiricism is, of course, a very old tradition. Back in the 1950s and 1960s, long before the speech work of the 1970s and 1980s, there was Skinner's Behaviorism in Psychology, Shannon's Information Theory in Electrical Engineering, and Harris' Distributional Hypothesis in American Linguistics and the Firthian approach in British Linguistics ("You shall know a word by the company it keeps"). It is possible that much of this work was actually inspired by Turing's code-breaking efforts during World War II, but we may never know for sure given the necessity for secrecy.

The recent revival in empiricism has been fueled by three developments. First computers are much more powerful and more available than they were in the 1950s when empiricist ideas were first applied to problems in language, or in the 1970s and 1980s, when data-intensive methods were too expensive for researchers working in universities. Second, data have become much more available than ever before. As a result of a number of data collection and related efforts such as ACL/DCI, BNC, CLR, ECI, EDR, LDC, ICAME, NERC, and TEI, most researchers should now be able to make use of a number of very respectable machine-readable dictionaries (MRDs) and text corpora. (See footnote 1 for information on contacting many of these organizations.) Data-intensive methods are no longer restricted to those working in affluent industrial laboratories. Third, and perhaps most importantly, due to various political and economic changes around the world, there is a greater emphasis these days on deliverables and evaluation. Data collection efforts have been relatively successful in responding to these pressures by delivering massive quantities of data. Text Analysis has also prospered because of its tradition of evaluating performance with theoretically motivated numerical measures such as entropy.

References

Ayuso, D.; Bobrow, R.; MacLaughlin, D.; Meteer, M.; Ramshaw, L.; Schwartz, R.; and Weischedel, R. (1990). "Toward understanding text with a very large vocabulary." *DARPA Speech and Natural Language Workshop*. San Mateo, CA, 354–358. Morgan Kaufmann.

Bahl, L.; Baker, J.; Cohen, P.; Dixon, N.; Jelinek, F.; Mercer, R.; and Silverman, H. (1975). "Preliminary results in the performance of a system for the automatic recognition of continuous speech." IBM Technical Report #RC 5654.

Bahl, L.; Baker, J.; Cohen, P.; Jelinek, F.; Lewis, B.; and Mercer, R. (1978). "Recognition of a continuously read natural corpus." In *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Bahl, L., and Mercer, R. (1976). "Part of speech assignment by a statistical decision algorithm." In *Abstracts of Papers from the International Symposium on Information Theory*.

Baker, J. (1979). "Trainable grammars for speech recognition." In *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, edited by Klatt and Wolf, 547–550.

Bakis, R. (1976). "Continuous speech recognition via centisecond acoustic states." In *Proceedings of 91st Meeting of the Acoustic Society of America*.

Baum, L. (1972). "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process." *Inequalities*, **3**, 1–8.

Bell, T.; Cleary, J.; and Witten, I. (1990). *Text Compression*. Prentice Hall.

Biber, D. (1993). "Representativeness in corpus design." *Computational Linguistics*, **19**(2). In press.

Boggess, L.; Agarwal, R.; and Davis, R. (1991). "Disambiguation of prepositional phrases in automatically labelled technical text." *AAAI*, 155–159.

Boguraev, B. (1990). "Looking Up: an account of the COBUILD project in lexical computing." *Computational Linguistics*, **16**(3), 184–185.

Brent, M. (1993). "Robust acquisition of subcategorization features from arbitrary text: syntactic knowledge meets unsupervised learning." *Computational Linguistics*, **19**(2). In press.

Brown, P.; Cocke, J.; Della Pietra, S.; Della Pietra, V.; Jelinek, F.; Lafferty, J.; Mercer, R.; and Rossin, P. (1990). "A statistical approach to machine translation." *Computational Linguistics*, **16**(2), 79–85.

Brown, P.; Della Pietra, S.; Della Pietra, V.; Lai, J.; and Mercer, R. (1992). "An estimate of an upper bound for the entropy of English." *Computational Linguistics*, **18**(1), 31–40.

Brown, P.; Della Pietra, S.; Della Pietra, V.; Mercer, R. (1993). "The mathematics of machine translation." *Computational Linguistics*, **19**(2). In press.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Church, K. (1988). "A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, Second Conference on Applied Natural Language Processing*. (ACL), Austin, TX, 136–143.

Church, K.; Hanks, P.; Hindle, D.; and Gale, W. (1991). "Using statistics in lexical analysis." In *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, edited by Zernik. Lawrence Erlbaum, 115–164.

Church, K., and Gale, W. (1991). "Concordances for parallel text." In *Proceedings, Seventh Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*.

Cohen, P., and Mercer, R. (1974). "The phonological component of an automatic speech-recognition system." In *Speech Recognition*, Invited Papers Presented at the 1974 IEEE Symposium, edited by R. Reddy, 275–320.

Cover, T., and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons.

Croft, W.; Turtle, H.; and Lewis, D. (1991). "The use of phrases and structured queries in information retrieval." In *SIGIR Forum*, edited by A. Bookstein, Y. Chiaramella, G. Salton, and V. Raghavan, 32–45.

Damerau, F. (1964). "A technique for computer detection and correction of spelling errors." *Communications of the ACM*, **7**(3), 171–176.

deMarcken, C. (1990). "Parsing the LOB corpus." Association for Computational Linguistics, 243–251.

DeRose, S. (1988). "Grammatical category disambiguation by statistical optimization." *Computational Linguistics,* **14**(1), 31–39.

Deroualt, A., and Merialdo, B. (1986). "Natural language modeling for phoneme-to-text transcription." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8(6), 742–749.

Fano, R. (1961). *Transmission of Information.* MIT. Press.

Firth, J. (1957). "A synopsis of linguistic theory 1930–1955." In *Studies in Linguistic Analysis,* Philological Society, Oxford; reprinted in *Selected Papers of J. R. Firth,* edited by F. Palmer. Longman. 1968.

Ford, M.; Bresnan, J.; and Kaplan, R. (1982). "A competence based theory of syntactic closure." In *The Mental Representation of Grammatical Relations,* edited by J. Bresnan. MIT Press, 727–796.

Frakes, W., and Baeza-Yates, R., eds. (1992). *Information Retrieval: Data Structures and Algorithms.* Prentice Hall.

Francis, W., and Kučera, H. (1982). *Frequency Analysis of English Usage,* Houghton Mifflin.

Furui, S. (1989). *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker.

Garside, R.; Leech, G.; and Sampson, G. (1987). *The Computational Analysis of English.* Longman.

Good, I. (1953). "The population frequencies of species and the estimation of population parameters." *Biometrika,* **40**, 237–264.

Halliday, M. (1966). "Lexis as a linguistic level," In *In Memory of J. R. Firth,* edited by C. Bazell, J. Catford, M. Halliday, and R. Robins. Longman.

Hanks, P. (1990). "Evidence and intuition in lexicography." In *Meaning and Lexicography,* edited by J. Tomaszczyk and B. Lewandowska-Tomaszczyk, 31–41. John Benjamins Publishing Company.

Hearst, M. (1991). "Toward noun homonym disambiguation using local context in large text corpora." In *Proceedings, Seventh Annual Conference of the UW Centre for the New OED and Text Research.* University of Waterloo, Waterloo, Ontario, 1–22.

Henisz-Dostert, B.; Ross Macdonald, R.; and Zarechnak, M., eds. (1979). *Machine Translation.* Mouton.

Hindle, D. (1989). "Acquiring disambiguation rules from text." ACL, 118–125.

Isabelle, P. (1992). "Bi-textual aids for translators." In *Proceedings, Eighth Annual Conference of the UW Centre for the New OED and Text Research.* University of Waterloo, Waterloo, Ontario, 76–89.

Jelinek, F. (1985). "Self-organized language modeling for speech recognition." IBM Report. Reprinted in *W & L,* 450–506.

Jelinek, F., and Mercer, R. (1980). "Interpolated estimation of Markov source parameters from sparse data." In *Proceedings, Workshop on Pattern Recognition in Practice.* North-Holland.

Jelinek, F., and Mercer, R. (1985). "Probability distribution estimation from sparse data." *IBM Technical Disclosure Bulletin,* 28, 2591–2594.

Jelinek, F.; Mercer, R.; and Roukos, S. (1991). "Principles of lexical language modeling for speech recognition." In *Advances in Speech Signal Processing,* edited by S. Furui and M. Mohan. Marcel Dekker, 651–700.

Karlsson, F. (1990). "Constraint grammar as a framework for parsing running text." In *Proceedings, 15th International Conference on Computational Linguistics (COLING-90),* 168–173.

Klavans, J., and Tzoukermann, E. (1990a). "The BICORD system." In *Proceedings, 15th International Conference on Computational Linguistics (COLING-90),* Helsinki, Finland, 174–179.

Klavans, J., and Tzoukermann, E. (1990b). "Linking bilingual corpora and machine readable dictionaries with the BICORD system." In *Proceedings, Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research,* 19–30.

Klatt, D. (1977). "Review of the ARPA speech understanding project." *Journal of the Acoustical Society of America.* Reprinted in *W & L,* 554–575.

Klatt, D. (1980). "Scriber and lafs: Two new approaches to speech analysis." In *Trends in Speech Recognition,* edited by W. Lea. Prentice-Hall.

Kukich, K. (1992). "Techniques for automatically correcting words in text." *ACM Computing Surveys,* **24**(4), 377–439.

Kupiec, J. (1989). "Augmenting a hidden Markov model for phrase-dependent word tagging." *DARPA Speech and Natural Language Workshop,* San Mateo, CA, 92–98. Morgan Kaufmann.

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

Lari, K., and Young, S. (1991). "Applications of stochastic context-free grammars using the inside-outside algorithm." *Computer Speech and Language,* 237–258.

Leech, G.; Garside, R.; and Atwell, E. (1983). "The automatic grammatical tagging of the LOB corpus." *ICAME News* **7**, 13–33.

Leonard, R. (1984). "A database for speaker-independent digit recognition." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **3**, 1–4.

Liberman, M., and Church, K. (1991). "Text analysis and word pronunciation in text-to-speech synthesis." In *Advances in Speech Signal Processing*, edited by S. Furui and M. Mohan. Marcel Dekker, 791–832.

Merialdo, B. (1991). "Tagging text with a probabilistic model." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 809–812.

Meyer, D.; Schvaneveldt, R.; and Ruddy, M. (1975). "Loci of contextual effects on visual word-recognition." In *Attention and Performance V*, edited by P. Rabbitt and S. Dornie. Academic Press, 98–116.

Minsky, M., and Papert, S. (1969). *Perceptrons; An Introduction to Computational Geometry*. MIT Press.

Mish, F., ed. (1983). *Webster's Ninth New Collegiate Dictionary*. Merriam, Webster.

Mosteller, F., and Tukey, J. (1977). *Data Analysis and Regression*. Addison-Wesley.

Mosteller, Fredrick, and Wallace, David (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Murray, K. (1977). *Caught in the Web of Words: James Murray and the Oxford English Dictionary*. Yale University Press.

Palermo, D., and Jenkins, J. (1964). *Word Association Norms*. University of Minnesota Press.

Price, P.; Fisher, W.; Bernstein, J.; and Pallett, D. (1988). "The DARPA 1000-word resource management database for continuous speech recognition." In *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **1**, 651–654.

Pustejovsky, J.; Berger, S.; and Anick, P. (1993). "Lexical semantic techniques for corpus analysis." *Computational Linguistics*, **19**(2). In press.

Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." In *Proceedings, IEEE*, **77**(2), 257–286. Reprinted in *W & L*, 267–296.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley.

Salton, G.; Zhao, Z.; and Buckley, C. (1990). "A simple syntactic approach for the generation of indexing phrases." Technical Report 90-1137, Department of Computer Science, Cornell University.

Sato, S., and Nagao, M. (1990). "Towards memory based translation." In *Proceedings, 15th International Conference on Computational Linguistics (COLING-90)*, 247–252.

Shannon, C. (1948). "The mathematical theory of communication." *Bell System Technical Journal*, **27**, 398–403.

Shannon, C. (1951). "Prediction and entropy of printed English." *Bell Systems Technical Journal*, **30**, 50–64.

Sinclair, J.; Hanks, P.; Fox, G.; Moon, R.; and Stock, P., eds. (1987). *Collins COBUILD English Language Dictionary*. Collins.

Sinclair, J., ed. (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Voutilainen, A.; Heikkila, J.; and Anttila, A. (1992). "Constraint grammar of English: A performance-oriented introduction." Publication No. 21, University of Helsinki, Department of Linguistics, Helsinki, Finland.

Waibel, A., and Lee, K., eds. (1990). *Readings in Speech Recognition*. Morgan Kaufmann.

Warwick-Armstrong, S., and Russell, G. (1990). "Bilingual concordancing and bilingual lexicography." Euralex 1990.

Weaver, W. (1949). "Translation." Reproduced in *Machine Translation of Languages*, edited in 1955 by W. Locke and A. Booth. MIT Press, 15–23.

Welch, T. (1984). "A technique for high performance data compression." *Computer*, **17**(6), 8–19.

Woods, W. (1970). "Transition networks for natural language analysis." *CACM*, **13**(10), 591–606.