Using Statistics in Lexical Analysis

Kenneth Church
William Gale
Patrick Hanks
Donald Hindle

Bell Laboratories and Oxford University Press

1. Introduction

The computational tools available for studying machine-readable corpora are at present still rather primitive. In the more advanced lexicographic organizations, there are *concordancing* programs (see figure below), which are basically KWIC (key word in context (Aho et al., 1988, p. 122), (Salton, 1989, p. 384)) indexes with additional features such as the ability to extend the context, sort leftwards as well as rightwards, and so on. There is very little interactive software. The lack of interactive software is perhaps part of the reason why dictionaries produced in the United States pay little attention to machine-readable corpora, and are based on collections of selected citations, augmented by introspection, rather than analysis of whole texts. The situation is somewhat different in Britain. British lexicographers, especially those working on dictionaries for foreign learners, are beginning to depend heavily on machine-readable corpora. They use these corpora and the basic concordancing tool mentioned above to fill in detailed syntactic descriptions (prompting a move, that will probably dominate lexicography in the 1990s, towards more thorough descriptions of lexical syntax). In the Cobuild project of the 1980s, for example, the typical procedure was that a lexicographer was given the concordances for a word or group of words, marked up the printout with colored pens in order to identify the salient senses, and then wrote syntactic descriptions and definitions.

Although this technology is an advance on using human readers to collect boxes of citation index cards (the method Murray used in constructing the Oxford English Dictionary a century ago, and still in use in some present-day lexicographic organizations), it works well only if there are no more than a few dozen concordance lines for a word, and just two or three main sense divisions. In analyzing complex words such as *strong* and *that*, the lexicographer is trying to pick out significant patterns and subtle distinctions that are buried in literally thousands of concordance lines. The unaided human mind simply cannot discover all the significant patterns, let alone group them and rank them in order of importance. Concordance analysis is still extremely labor-intensive, and prone to errors of omission.

There are similar problems in Information Retrieval (Salton, 1989). Keyword systems work best when there are only a few dozen hits. But unfortunately, it is very easy for a user to select a keyword like *food*, and be buried under thousands of documents. There ought to be a set of tools that make it easier for a user to cope with a very large set of documents. In particular, the user should be able to ask the system to suggest a set of candidate keywords that would help disambiguate among the various senses of *food*, so that he can quickly focus on the sense that he[1] is interested in.

Computational linguists run into similar problems when they try to write grammars, especially

1. The reader is asked to interpret our use of *he* (and subsequent uses of *she*) as gender-neutral, given the unfortunate lexical gap in English in this respect.

A very small sample of the concordances to ‘‘strong’’ (from 1988 AP newswire)

f somebody catching it has become quite	strong	, ’ ’’ the newspaper said . *E* *S* The Monitor said
necessarily appear on the surface to be	strong	, ’’ said McGovern , who first drew attention in the
the actress . *E* *S* Kristy is ‘‘ very	strong	, although she doesn’t necessarily appear on the surf
eratures . *E* *S* ‘‘ What we need is a	strong	, energetic , young , brilliant man , and that’s what
S* ‘‘ You know , the Soviet Union has a	strong	, energetic man , ’’ Cash told about 150 people who s
rt showed . *E* *S* The impression of a	strong	, potentially inflationary economy was heightened by
or the November election . *E* *S* ‘‘ A	strong	, well-financed Republican Party will be a benefit to
mathematics is regarded in the West as	strong	. *E* *S* It is not known exactly what changes the Ce
evious months . *E* *S* Sales were up a	strong	1.2 percent in December and 0.3 percent in November ,
about Mr . Gorbachev and they welcomed	strong	American leadership of the NATO alliance . *E* *S* We
et Ambassador Yuri Dubinin to receive a	strong	U.S . protest and that Defense Secretary Frank C . Ca
uded Hughes ’ direction . *E* *S* ‘‘ As	strong	and independent as I come off on the set , I need a d
rtner . *E* *S* Our commercial ties are	strong	and of great benefit to people on both sides of the b
analyst Linda Simard said crude opened	strong	at the start , picking up on moderate overnight gains
f follow-through buying from Thursday’s	strong	close . *E* *S* Early trading volume was light ahead
is energetic person-to-person style and	strong	conservative message will make him the conservative a
c population , always have maintained a	strong	cultural and ethnic identity . *E* *S* One of the Est
themselves ... and we’ve got to have a	strong	defense . ’’ .End of Discourse *E* *S* .Story 88 /mur
0457 *E* *S* TAIPEI , Taiwan (AP) - A	strong	earthquake centered off Taiwan’s eastern coast violen
s , some analysts said the figures were	strong	enough to indicate consumers were not dragging the ec
s pointed toward the December report as	strong	evidence of the long-awaited reversal in the nation’s
5.8 billion Canadian dollars largely on	strong	foreign sales of forest products . *E* *S* However ,
, and basically a black school that was	strong	in academics , ’’ Dade said . *E* *S* ‘‘ Before , we
finishing third in Iowa , maintained a	strong	lead in New Hampshire - but he no longer had the huge
etts Gov . Michael Dukakis maintained a	strong	lead in the Democratic race . *E* *S* ABC reported he
S* In both polls , Dukakis maintained a	strong	lead in the Democratic race . .End of Discourse *E* *
er whose poll you’re looking at - and a	strong	one , too , ’’ said Jeff Alderman , chief of polling
port on the seacoast . *E* *S* Kemp , a	strong	proponent of states ’ rights , has asked federal regu
rsuit of peace , NATO must soon offer a	strong	proposal on conventional and chemical weapons control
rsuit of peace , NATO must soon offer a	strong	proposal on conventional and chemical weapons control
ri Dubinin Friday morning to ‘‘ lodge a	strong	protest . *E* *S* ’’ Defense Secretary Frank C . Carl
er Alexander Bessmertnykh read him a ‘‘	strong	protest . *E* *S* ‘‘ The Soviet side cannot but view
the administration immediately lodged a	strong	protest with the Soviet ambassador here , saying the
her ; its contributors are regarded as	strong	researchers and theorists , he said . *E* *S* Sternbe
nder seven presidents and recalling his	strong	role as chief of staff in the White House during the
ock . *E* *S* ‘‘ We wanted a positive ,	strong	school , and basically a black school that was strong
to racial problems by failing to take a	strong	stand on civil rights . *E* *S* ‘‘ Supremacists seem
to racial problems by failing to take a	strong	stand on civil rights . *E* *S* ‘‘ Supremacists seem
ship . *E* *S* Those are but two of his	strong	suits . *E* *S* ’’ The newspaper has a circulation of
more than a foot of snow and unleashing	strong	wind as schools , offices and roads were closed . *E*

A very small sample of the concordances to ‘‘powerful’’ (from 1988 AP newswire)

e family , said to be the nation’s most powerful	, and even shot a labor leader as a favor to repute
S* It’s fair taxes for the rich and the powerful	. ’’ .PP *E* *S* .End of Discourse *E* *S* .Story 8
illumination rounds ’’ from the ship’s powerful	5-inch gun . *E* *S* As soon as the flares went off
izumi , a city built around 1100 by the powerful	Fujiwara clan that was once a center of political p
ry officials had become involved with a powerful	Honduras-based narcotics trafficker , Juan Ramon Ma
Nikolai V . Talyzin as chairman of the powerful	State Planning Committee , deputy chairman Council
Nikolai V . Talyzin as chairman of the powerful	State Planning Committee , deputy chairman Council
s been rejected by the largest and most powerful	Tamil militant group , the Liberation Tigers of Tam
ion where drug traffickers can become a powerful	and disruptive factor in a society . *E* *S* Colomb
rer called ‘‘ Empire of the Sun ’’ ‘‘ a powerful	and richly human anti-war film . *E* *S* ’’ Spielbe
ed an explosive punch up to 11 times as powerful	as the atomic bomb that devastated Hiroshima . *E*
8/02/18/a0520 *E* *S* BOSTON (AP) - A powerful	blast of shock waves can smash gallstones inside th
ted . *E* *S* It also said two other ‘‘ powerful	bombs ’’ were defused ‘‘ in the last several days ’
ederation of Economic Organizations , a powerful	business alliance , is planning a leap into the 21s
itian army Col . Jean-Claude Paul , the powerful	commander of the key batallion in Port-au-Prince ,
. *E* *S* Despite the existence of two powerful	drugs to treat the rare form of pneumonia , scienti
and simulated windsurfing in front of a powerful	fan . *E* *S* Among the people wearing shorts were
nd West Germany , both with politically powerful	farming lobbies , have sought an increase of $3.1 b
till was a land of barbarian tribes and powerful	feudal warriors - one of Japan’s last frontiers . *
out . *E* *S* ‘‘ It’s a very silent but powerful	force in Southern politics , ’’ Rose said . *E* *S*
en . *E* *S* The reflex is particularly powerful	in children , doctors say . *E* *S* Kendall was in
en . *E* *S* The reflex is particularly powerful	in children , doctors say . *E* *S* Tecklenburg sai
ficient in the short-term , it provides powerful	incentive for workers to sabotage innovative techno
eighth straight term . *E* *S* With the powerful	infrastructure of the governing Colorado Party at h
k was retained as head of South Korea’s powerful	intelligence agency , the Agency for National Secur
hn Moo-hyuk was retained as head of the powerful	intellignece organization , the Agency for National
industry’s steady transition to a more powerful	internal ‘‘ traffic cop . *E* *S* ’’ Ashton-Tate Co
Ermann said Reagan ‘‘ is sending a very powerful	message to many millions of people all over the wor
Oh is expected to be Roh’s link to the powerful	military , which has long dominated the country . *
E* *S* The Khmer Rouge remains the most powerful	military force within the coalition . *E* *S* For m
idency in 1985 but still holds the most powerful	political position in the East Africa country as le
nt party is considered by many the most powerful	post in Italy and some party insiders would like to
icy . *E* *S* Razumovsky also holds the powerful	post of secretary of the Central Committee in charg
mental device would be the world’s most powerful	proton collider , capable of conducting research in
miles long , would be the world’s most powerful	proton collider . *E* *S* Its goal is fundamental r
. *E* *S* But (heavy metal) is a very powerful	reinforcement . *E* *S* It legitimizes the nasty st
is huge there . *E* *S* ’’ It is also a powerful	source of radiation , emitting infrared , X-rays an
Boesky , once one of Wall Street’s most powerful	stock speculators . *E* *S* He paid a record $100 m
troduced tough measure to curb the once powerful	unions and the last major industrial strike in Brit
E* *S* Women have become a distinct and powerful	voting group in recent years , Ms . Ferraro said ,

disambiguation rules to diagnose lexical ambiguity. Again, the most common words cause the most trouble. Consider the word *that*, which has many parts of speech, and is frequently found in expressions such as: *that way, that moment, that point, so that, think that, now that, indicated that, all that*, etc. Like concordance analysis, designing disambiguation rules is still extremely labor-intensive, and prone to errors of omission. The unaided grammar writer simply cannot discover all the significant patterns, let alone group them and rank them in order of importance.

If the tools were better, computational linguists could attempt to model many more sources of constraint than they are able to deal with right now. For example, a parser really ought to be able to take advantage of the fact that *eating food* and *drinking water* are much more plausible that *eating water* and *drinking food*, but it is currently just too labor-intensive to deal with facts such as these.

This paper will discuss a number of statistical tools and show some examples of how these tools can be used to enhance productivity of a human trying to solve one of these problems. It will be assumed that we can depend on human judgment to use the statistics appropriately, and to check the results to see that they are reasonable. The emphasis on human interaction distinguishes our approach from self-organizing approaches such as Jelinek (1985).

Specifically, we will discuss three steps requiring human judgment:

1. Choose an appropriate statistic (e.g., mutual information, t-score),
2. preprocess the corpus to highlight properties of interest (with a part of speech tagger or a parser), and
3. select an appropriate unit of text (e.g., bigram, SVO triple, discourse).

We will begin the discussion by introducing two statistics, mutual information and the t-score, and demonstrate that they answer different questions. Given an array of statistical tools that can be used to some a wide range of tasks in lexicography and information retrieval, users will, we believe, rapidly learn to select the ones that are most appropproriate for their particular task.

2. Step 1: Select Appropriate Statistic

2.1 Mutual Information: A Measure of Similarity

Church and Hanks (1989) discussed the use of the mutual information statistic in order to identify a variety of interesting linguistic phenomena, ranging from semantic relations of the doctor/nurse type (content word/content word) to lexico-syntactic co-occurrence preferences between verbs and prepositions (content word/function word).[2] Mutual information, $I(x;y)$, compares the probability of observing word x and word y *together* (the joint probability) with the probabilities of observing x and y *independently* (chance).

$$I(x;y) \equiv \log_2 \frac{P(x,y)}{P(x)\ P(y)}$$

If there is a genuine association between x and y, then the joint probability $P(x,y)$ will be much larger than chance $P(x)\ P(y)$, and consequently $I(x;y) >> 0$, as illustrated in table 1 (below). If there is no interesting relationship between x and y, then $P(x,y) \approx P(x)\ P(y)$, and thus, $I(x;y) \approx 0$. If x and y are in

2. This statistic is also discussed by Jelinek (1985) for applications in speech recognition.

complementary distribution, then $P(x,y)$ will be much less than $P(x)\ P(y)$, forcing $I(x;y) << 0$. Word probabilities, $P(x)$ and $P(y)$, are estimated by counting the number of observations of x and y in a corpus, $f(x)$ and $f(y)$, and normalizing by N, the size of the corpus. Joint probabilities, $P(x,y)$, are estimated by counting the number of times that x is followed by y, $f(x,y)$, and normalizing by N.

Table 1 shows the mutual information and frequency values for twenty pairs of words. The frequency values were computed over our corpus of 1988 Associated Press newswire (N = 44.3 million words). In particular, the table shows that $I(strong;northerly)$ has a mutual information value of 10.47 because $\log_2((7\times N)/(7809\times 28)) = 10.47$. The table shows the top ten scoring pairs of the form *strong* ____, and the top ten scoring pairs of the form *powerful* ___.

Note that pairs with very high mutual information values are generally quite strongly associated. In (Church and Hanks, 1989), we argued that a table of mutual information values such as these could be used as an index to a concordance. Mutual information can help us decide what to look for in the concordance; it provides a quick summary of what company our words do keep (Firth, 1957).

Table 1: Some Interesting Associations with *strong* and *powerful*
in the 1988 AP Corpus (N = 44.3 million)

I(x;y)	fxy	fx	fy	x	y
10.47	7	7809	28	strong	northerly
9.76	23	7809	151	strong	showings
9.30	7	7809	63	strong	believer
9.22	14	7809	133	strong	second-place
9.17	6	7809	59	strong	runup
9.04	10	7809	108	strong	currents
8.85	62	7809	762	strong	supporter
8.84	8	7809	99	strong	proponent
8.68	15	7809	208	strong	thunderstorm
8.45	7	7809	114	strong	odor
8.66	7	1984	388	powerful	legacy
8.58	7	1984	410	powerful	tool
8.35	8	1984	548	powerful	storms
8.32	31	1984	2169	powerful	minority
8.14	9	1984	714	powerful	neighbor
7.98	9	1984	794	powerful	Tamil
7.93	8	1984	734	powerful	symbol
7.74	32	1984	3336	powerful	figure
7.54	10	1984	1204	powerful	weapon
7.47	24	1984	3029	powerful	post

Strong and *powerful* are an interesting pair to compare and contrast because their meanings are so similar. The connection between the two words can be found in any thesaurus including *Roget's International Thesaurus, Fourth Edition* (Chapman, 1977, paragraphs 157.12, 159.13) and *Chambers 20th Century Thesaurus* (Seaton et al., 1986, p. 482, 606). It is also not very hard to discover that there is a relation between *strong* and *powerful* from a machine readable dictionary. In the *Cobuild Dictionary* (Sinclair et al., 1987), for example, the definition of *strong* contains 5 references to the word *powerful*, and the definition of *powerful* contains 8 references to the word *strong*. In addition, there are also a number of references to morphological variants such as *strongly* and *powerfully*.

2.2 t-test: A Measure of Dissimilarity

Although it is easy to see the similarity between *strong* and *powerful* after consulting these references, it is more difficult to see the difference. For example, given the fragment from the *Oxford Advanced Learner's Dictionary of Current English, Fourth Edition*, henceforth OALDCE4:

> *strong* (OALDCE4):
> **4(a)**(capable of) having a great effect on the senses; intense or powerful; *a strong light, colour; a strong feeling of nausea; Her breath is rather strong,* ie has an unpleasant smell. (Cowie, 1989, definition of *strong*, p. 1276)

a reader might conclude that *Her breath is rather strong* has more or less the same meaning as *Her breath is rather powerful*. George Miller (personal communication) has observed that school children often misuse dictionaries in just this way, and inappropriately substitute nearly synonymous words into the example sentences. Dictionaries, especially learner's dictionaries, are very good at identifying related words, but they don't always succeed in describing the subtle distinctions among related words. Perhaps they could do a better job in describing these distinctions if they had access to better tools.

Smadja (1989) has argued basically the same point, using Halliday's observation that the collocation *strong tea* is much more plausible than *powerful tea*.

> ''The fact that people prefer saying *drink strong tea* to *powerful tea*, and prefer saying *drive a powerful car* to *a strong car* cannot be accounted for on pure syntactic or semantic grounds. These are lexical constraints that need to be introduced in order to filter out such oddities when producing English... Such lexical relations represent *idiosyncratic collocations* and account for a large part of English word combinations... They need to be specifically included in dictionaries... For language generation, this type of lexical knowledge is crucial to the problem of lexical choice... To bring co-occurrence knowledge to bear in language generation, there is a need for... automatically extracted co-occurrence knowledge...'' (Smadja, 1989)

Halliday put the argument this way:

> ''*...a strong car* and *powerful tea* will either be rejected as ungrammatical (or unlexical) or shown to be in some sort of marked contrast with *a powerful car* and *strong tea*; in either case the paradigmatic relation of *strong* to *powerful* is not a constant but depends on the syntagmatic relation into which each enters, here with argument *car* or *tea*.'' (Halliday, 1966, p. 150)

Two pages later, Halliday proposes an approach for identifying collocations, which is similar in spirit to the mutual information tool that we have been discussing:

> In place of the highly abstract relation of structure, ... lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either a scale or at least a cut-off point.'' (Halliday, 1966, p. 152)

However, when we look at the mutual information statistic in more detail, we find that it is probably not the most appropriate way to establish differences among nearly synonymous words such as *strong* and

powerful. Although mutual information is an extremely useful statistic, it is based on certain assumptions which have their limitations. In particular, it is difficult to make negative statements. Consider the hypothesis that the collocation *strong support* is much more plausible than *powerful support*. (We would have liked to use Halliday's example of *strong tea* and *powerful tea* here, but we use the *strong support* example instead because the AP has much more discussion about politics than about *tea*.)[3] Suppose we wanted to find evidence to account for the difference between *strong support* and *powerful support*. We will quickly find that it is much easier to find evidence for *strong support* than to find the lack of evidence for *powerful support*. We must be careful not to fall into the failure-to-find fallacy. That is, when you don't have much evidence for something, it is very hard to know whether it is because it doesn't happen, or because you haven't been looking for it in the right way (or in the right place).

We would be able to say that *powerful support* is implausible if we could establish that $I(\textit{powerful};\textit{support}) << 0$. However, we are rarely able to observe mutual information scores much less than zero because our corpora are too small (and our measurement techniques are too crude). Suppose, for example, that two words, x and y, both appear about 10 times per million words of text. Then, $P(x) = P(y) = 10^{-5}$ and chance is $P(x)\ P(y) = 10^{-10}$. Thus, to say that $I(x,y)$ is much less than zero, we need to say that $P(x,y)$ is much less than 10^{-10}, a statement that is hard to make with much confidence given the size of presently available corpora. In fact, we cannot (easily) observe a probability less than $1/N \approx 10^{-7}$, and therefore it is hard to know if $I(x,y)$ is much less than chance or not, unless chance is very large.

However, it is possible to rephrase the question so that we can obtain a usable result. Instead of asking what doesn't happen after *powerful*, let's ask which words are significantly more likely to appear after *strong* than after *powerful* (in AP journalese, at least). In this way, we can show that *strong support* is significantly more likely than *powerful support*, and thus we are able to make a negative statement about *powerful support* (relatively speaking). Note that we couldn't make an absolute statement because we don't have enough evidence to say that *powerful support* is less likely than chance. In fact, what little evidence we have seems to suggest just the opposite. That is, the mutual information of *powerful support* is positive (1.74),[4] which means that its probability is approximately three times[5] greater than chance. Of course, the variances are also quite large, so that a t-score would not be significant:

---

3. *support* is about twice as common in the 1988 AP corpus (13,428 instances in 44.3 million words $\approx$ 300 references per million words) as in the Brown Corpus (177 references in 1 million words). In contrast, *tea* is almost four times less common in the 1988 AP corpus; there were 322 references to *tea* in the 1988 AP corpus (7.3 references per million words), whereas there were 27 references in the Brown Corpus.

4. $I(\textit{powerful};\ \textit{support})$ is computed to be 1.74, by the following calculation:

$$\log_2 \frac{f(\textit{powerful support}) \times N}{f(\textit{powerful}) \times f(\textit{support})} \approx \log_2 \frac{2 \times N}{1984 \times 13{,}428} \approx 1.74$$

with the following values: $N = 44.3$ million; $f(\textit{powerful support}) = 2$; $f(\textit{powerful}) = 1984$; $f(\textit{support}) = 13{,}428$.

5. $2^{1.74} \approx 3$.

$$t = \frac{P(\textit{powerful support}) - P(\textit{powerful})\ P(\textit{support})}{\sqrt{\sigma^2(P(\textit{powerful support})) + \sigma^2(P(\textit{powerful})\ P(\textit{support}))}}$$

$$< \frac{\frac{f(\textit{powerful support})}{N} - \frac{f(\textit{powerful})\ f(\textit{support})}{N^2}}{\frac{\sqrt{f(\textit{powerful support})}}{N}}$$

$$\approx \frac{2 - \frac{1984 \times 13{,}428}{N}}{\sqrt{2}} \approx 0.99$$

In other words, although $P(\textit{powerful support})$ is greater than $P(\textit{powerful})\ P(\textit{support})$, the difference is less than one standard deviation (σ), which isn't significant. Normally, we would want at least a difference of 1.65 standard deviations, so that we could have 95% confidence that the difference was real, and not due to chance. With a difference of only one standard deviation, there is about a 30% chance that the difference is a fluke. Thus, we do not get a usable result if we try to compare $P(\textit{powerful support})$ with $P(\textit{powerful})\ P(\textit{support})$. In contrast, if we compare $P(\textit{powerful support})$ with $P(\textit{strong support})$, the result is highly significant:

$$t = \frac{P(\textit{powerful support}) - P(\textit{strong support})}{\sqrt{\sigma^2(P(\textit{powerful support})) + \sigma^2(P(\textit{strong support}))}}$$

$$\approx \frac{\frac{f(\textit{powerful support})}{N} - \frac{f(\textit{strong support})}{N}}{\sqrt{\frac{f(\textit{powerful support})}{N^2} + \frac{f(\textit{strong support})}{N^2}}} \approx \frac{2 - 175}{\sqrt{2 + 175}} \approx -13$$

In other words, $P(\textit{powerful support})$ is thirteen standard deviations less likely than $P(\textit{strong support})$. We can very confidently reject the null hypothesis that there is no difference between the two.

Table 2 presents some results of using the t-score statistic to contrast *strong w* with *powerful w*. The left half of the table shows ten words that are much more likely to appear after *strong* than after *powerful*. The right half shows ten words that are more likely to appear after *powerful* than after *strong*. The t-scores were computed by the formula:[6]

$$t \equiv \frac{P(w|\textit{strong}) - P(w|\textit{powerful})}{\sqrt{\sigma^2(P(w|\textit{strong})) + \sigma^2(P(w|\textit{powerful}))}}$$

The t-score indicates the difference between $P(w|\textit{strong})$ and $P(w|\textit{powerful})$ in standard deviations. The probabilities, $P(w|\textit{strong})$ and $P(w|\textit{powerful})$ could be estimated by the maximum likelihood method

6. The vertical bar is to be read as specifying a conditional probability -- e.g. $P(w|\textit{strong})$ should be interpreted as the probability of a word *w* occurring given an occurrence of *strong* (in this case, as the previous word).

(MLE), which would simply divide $f(strong, w)$ and $f(powerful, w)$ by $f(strong)$ and $f(powerful)$, respectively. The variances, $\sigma^2 P(w|strong)$ and $\sigma^2 P(w|powerful)$ would be estimated by dividing $f(strong, w)$ and $f(powerful, w)$ by $f(strong)^2$ and $f(powerful)^2$, respectively. However, this method is seriously flawed when the counts are very small.

We should probably use the Good-Turing (GT) estimates (Good, 1953) instead of MLE, but we have decided to use a compromise so that the reader would have an easier time replicating our results. We simply add 1/2 to all frequency counts (and adjust the denominator appropriately so that $f(strong) = \sum_w f(strong, w)$ and $f(powerful) = \sum_w f(powerful, w)$). See Box and Tiao (1973) for a discussion of this method, which we call the ELE (Expected Likelihood Estimator). We have checked the t-scores computed by the ELE with those computed by the GT methods on many of the examples in this paper and found that the differences are acceptable for our purposes.[7] The order is nearly preserved, though the magnitude of the t-scores with ELE method are uniformly about 30% too large. Thus, the 1.65 threshold should probably be adjusted upwards to about 2.15. In any case, the t-scores in table 2 are all highly significant.

The second t-score in table 2 (11.94) compares *strong support* with *powerful support*. It is computed as follows (with $N = 44.3$ million; $f(strong\ support) = 2$; $f(powerful) = 1984$; $f(support) = 13{,}428$; $V = 1841$):[8]

$$t \equiv \frac{P(w|strong) - P(w|powerful)}{\sqrt{\sigma^2(P(w|strong)) + \sigma^2(P(w|powerful))}}$$

$$\approx \frac{\frac{f(strong\ support) + 1/2}{f(strong) + V/2} - \frac{f(powerful\ support) + 1/2}{f(powerful) + V/2}}{\sqrt{\frac{f(strong\ support) + 1/2}{(f(strong) + V/2)^2} + \frac{f(powerful\ support) + 1/2}{(f(powerful) + V/2)^2}}}$$

$$\approx \frac{\frac{175.5}{7809 + 1841/2} - \frac{2.5}{1984 + 1841/2}}{\sqrt{\frac{175.5}{(7809 + 1841/2)^2} + \frac{2.5}{(1984 + 1841/2)^2}}} \approx 11.94$$

7. In other applications, we have found that the ELE is much less acceptable (Gale and Church, 1990). It is clear that the ELE has many problems. In particular, there are much better methods for estimating the probability and variance of types that have not been seen. However, it happens that the t-score calculation is not very sensitive to the errors introduced by ELE, because the t-score tends to depend much more on the quantities with larger counts. In contrast, mutual information is much more sensitive to these errors since it depends very strongly on quantities with small counts.

8. V is the number of words that follow either *strong* or *powerful*. It is required by the ELE method so that the sum of the estimated probabilities will be one.

Table 2: An Example of the t-score

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

The following table is presented in order to illustrate that mutual information, a measure of similarity, is answering a different question than the t-test, a measure of dissimilarity. The first and last columns are repeated from table 1. The second, third and fourth columns show the t-scores, $f(strong, w)$ and $f(powerful, w)$, respectively. Note that it is possible for a word to have a high mutual information score and a low t-score. For example, *strong* and *thunderstorms* are highly associated, but we cannot say (with very much confidence) that *strong thunderstorms* is more likely than *powerful thunderstorms*. The difference between 20 and 4 is not significant because there are many more references to *strong* (7809) than to *powerful* (1984).

Table 3: Answer Different Questions

Associated with strong					Associated with powerful				
I(strong; w)	t	strong	powerful	w	I(powerful; w)	t	strong	powerful	w
10.47	1.73	7	0	northerly	8.66	–2.53	1	7	legacy
9.76	3.12	23	1	showings	8.58	–2.67	0	7	tool
9.30	1.73	7	0	believer	8.35	–2.33	4	8	storms
9.22	2.98	14	0	second-place	8.32	–5.37	3	31	minority
9.17	1.51	6	0	runup	8.14	–3.02	0	9	neighbor
9.04	1.22	10	1	currents	7.98	–3.02	0	9	Tamil
8.85	7.45	62	0	supporter	7.93	–2.59	2	8	symbol
8.84	1.94	8	0	proponent	7.74	–3.89	0	15	figures
8.68	0.89	20	4	thunderstorms	7.54	–3.18	0	10	weapon
8.45	1.73	7	0	odor	7.47	–4.91	0	24	post

How can a lexicographer make use of statistics of this kind? Two possibilities are immediately apparent. In the first place, they might encourage lexicographers to sharpen the focus of definitions, highlighting salient facts and omitting the remote possibilities that occur only to nervous lexicographers, anxious to cover all possible eventualities. In the second place, they might be used to formulate explicit rules for choosing among near synonyms. When is it better to talk about *strong support*, and when is *powerful support* more appropriate?

There is a long tradition of explicit synonym studies of precisely this kind in American Collegiate and Unabridged dictionaries. (British dictionaries, with few exceptions, do not contain studies of this kind.) It seems likely that the value of such studies could be enhanced if they are based on a selection of statistically

significant evidence to augment the insights and pontifications of the lexicographer.

A couple of the relevant definitions for *strong* and *powerful* in two American Unabridged dictionaries, *Merriam Webster's Third New International* (1961), henceforth MW3, and the *Random House Dictionary, Second Edition* (1987), henceforth RHD2, are given below. Also mentioned are the synonym studies, including one for *power* in MW3 which mentions *strength*.

strong (RHD2):
6. powerful in influence, authority, resources, or means of prevailing or succeeding: *a strong nation*.
7. compelling; of great force, effectiveness, potency, or cogency: *strong reasons, strong arguments*.

powerful (RHD2):
4. potent, efficacious: *a powerful drug*.
5. having great effectiveness, as a speech, speaker, description, reason, etc.
6. having great power, authority, or influence; mighty: *a powerful nation*.

Even though there is a synonym study for *powerful* in RHD2, *strong* is not one of the words studied. There is little indication in this dictionary of the difference between the two words, even though there is a synonym study at *powerful* (distinguishing *mighty* and *potent*).

strong (MW3):
3 having or exhibiting moral or intellectual force, endurance, or vigor <mistook an opinionated mind for a strong one...> <strong ruler> <strong president>.

A synonym study for *strong* in MW3 distinguishes *stout, sturdy, stalwart, tough,* and *tenacious*.

powerful (MW3):
1a having great force or potency: STRONG, COMPELLING ... **b** having great prestige or effect: INFLUENTIAL, STIMULATING.

There is a synonym study at *power* which mentions, inter alia, *strength*.

''POWER signifies ability, latent, exerted, physical, mental or spiritual, to act, be acted upon, effect, or be effected, sometimes designating the thing having this ability...

STRENGTH applies to the power residing in a thing as a result of qualities or properties (as health or soundness in bodily condition, or numbers or great equipment in military organization) that enable it to exert force or manifest great energy as in resistance, attack, or endurance....''

Although there are interesting hints and suggestions buried in these two dictionaries, they do not provide a sharp and clear criterion for distinguishing, say, *a strong nation* from *a powerful nation*. Factors that might

be relevant seem to be obscured by being buried in a welter of wording.

If we turn now to the statistically selected comparison, we get some support for a generalization about the nature of the distinction. A *strong nation* has something in common with *strong defense, strong economy*, and *strong growth*. A *powerful nation* has something in common with *powerful posts* (*military* and *political*), a *powerful figure*, and a *powerful presidency*.

An important criterion for differentiation seems to be that *strong* tends to denote an intrinsic quality, whereas *powerful* appears to be extrinsic, referring more to the effect on others or on the external world. Any worthwhile politician or cause can expect *strong supporters*, who are enthusiastic, convinced, vociferous, etc. But far more valuable are *powerful supporters*, who will brings others with them. They are also, according to the AP news, much rarer -- or at any rate, much less often mentioned.

Like many good lexicographic insights, if true, this may seem so blindingly obvious as to be hardly worth stating. It is worth reminding ourselves, therefore, that two great American dictionaries did not, apparently, have sufficient evidence to prompt this particular distinction. Availability of evidence of the kind described in this paper will help lexicographers to bring their subject into focus, word by word and collocation by collocation. Of course, we should beware of the dange of simplistic overextension of such criteria. For example, the criterion proposed here does not shed much light on the reasons why English speakers prefer to talk about *strong tea* and *strong liquor* rather than *powerful tea* and *powerful liquor*. At best we can use the criterion to say something about cultural attitudes to tea and liquor, and opposed, say, to (powerful) drugs. There is little in the real world that justifies the distinction. For purposes of lexical analysis, therefore, it is probably wisest to assume, with Halliday, that not all of the syntagmatic relations identified by the statistic will have a clear semantic motivation.

2.3 Scale Statistics

We believe the lexicographer should decide which statistic is more appropriate for his application. The choice, of course, will vary from case to case. Mutual information is more helpful in identifying associations (similarities) whereas the t-score focuses more on subtle distinctions (differences). In some cases, the lexicographer may want to use both statistics, as we did in the discussion just above.

Of course, mutual information and t-scores are not the only statistics to choose from. This section will introduce two more, the mean and variance of the separation between a pair of words.

In Church and Hanks (1989), we used table 4 (below) to demonstrate that different linguistic preferences operate at different scales. (The following argument was inspired by Smadja (1989)). In fixed expressions, such as *bread and butter* and *drink and drive*, the words of interest are separated by a fixed number of words and there is very little variance. In the 1988 AP, it was found that the two words are always exactly two words apart whenever they are found near each other (within five words). That is, the mean separation is two, and the variance is zero. Compounds also have very fixed word order (little variance), but the average separation is closer to one word rather than two. In contrast, relations such as *man/woman* are less fixed, as indicated by a larger variance in their separation. (The nearly zero value for the mean separation for *man/women* indicates that words appear about equally often in either order.) Lexical relations come in several varieties. There are some like *refraining from* which are fairly fixed, and others like *keeping* (*someone or something*) *from* which are almost certain to be separated by a direct object.

Table 4: Mean and Variance of the Separation Between X and Y

Relation	Word x	Word y	Separation	
			mean	variance
fixed	*bread*	*butter*	2.00	0.00
	drink	*drive*	2.00	0.00
compound	*computer*	*scientist*	1.12	0.10
	United	*States*	0.98	0.14
semantic	*man*	*woman*	1.46	8.07
	man	*women*	–0.12	13.08
lexical	*refraining*	*from*	1.11	0.20
	coming	*from*	0.83	2.89
	keeping	*from*	2.14	5.53

A lexicographer could use these two statistics to help separate the different types of relations. Thus, for example, he could use the mean separation and its variance to distinguish verb/preposition combinations that usually take a direct object (e.g., *keeping from*) from those that usually do not (e.g., *refraining from*).

In addition, table 4 shows quite clearly that various statistics computed over bigrams will pick out some interesting facts (those with a separation of 1), but will miss some others (those with larger separations). Although this problem could be fixed by extending the window size in some way to allow more separation, most such methods will then smear (defocus) some of the facts at smaller scales. It is probably necessary that the lexicographer adjust the window size to match the scale of the phenomena that he is interested in. We will return to this point when we discuss step 3, selecting the appropriate unit of text (bigram, clause, or discourse).

3. Step 2: Preprocessing the Corpus

3.1 Preprocessing with a Part of Speech Tagger

There are many ways in which the lexicographer might want to preprocess the corpus. We will discuss two here: (1) tagging each word with a part of speech, and (2) parsing each clause into an subject-verb-object (SVO) triple. (No preprocessing is, of course, another option.)

Let us first consider an application where a part of speech tagger can be of considerable value: designing disambiguation rules for *to* and *that*. Let us start by considering the words immediately preceding *to*. Which of these words can be used to decide that the *to* is an infinitive marker and which of these can be used to decide that the *to* is a preposition?

We will use the same t-score argument as before, but this time, we will use the Tagged Brown Corpus (Francis and Kucera, 1982) instead of the 1988 AP corpus. Table 5 (below) shows a small sample of the results. The contrast is fairly compelling. The words on the left side of table 5 are strong indicators that the *to* is an infinitive marker, whereas the words on the right side of the table are strong indicators that the *to* is a preposition.

A lexicographer would also be interested in being able to distinguish verbs that take infinitival complements from those that take prepositional complements. Many current dictionaries, especially learner's dictionaries such as *Cobuild* (Sinclair et al., 1987) and the *Oxford Advanced Learners Dictionary* (Cowie, 1989) are supposed to describe the complements of common verbs, but there are still gaps in these dictionaries. The t-score would be very handy for identifying these gaps and helping the lexicographer to improve the coverage of complement structures in systematic ways.

Table 5: Which words precede the infinitival use of *to* (to/to) and which words precede the prepositional use of *to* (to/in)?

Infinitival use of to				Prepositional use of to			
t	w to/to	w to/in	w	t	w to/to	w to/in	w
16.01	266	2	had/hvd	–12.44	10	176	back/rb
15.58	268	6	have/hv	–9.92	0	99	according/in
13.60	245	16	is/bez	–9.50	9	109	went/vbd
13.58	190	1	able/jj	–8.90	7	94	go/vb
12.59	160	0	want/vb	–8.54	29	125	up/rp
12.08	188	11	was/bedz	–8.38	3	77	as/in
11.77	140	0	began/vbd	–8.08	1	68	respect/nn
11.37	135	1	trying/vbg	–7.64	1	61	addition/nn
10.25	122	4	order/nn	–7.63	14	85	down/rp
10.07	107	1	wanted/vbd	–7.57	1	60	close/rb
9.86	202	34	going/vbg	–7.17	0	52	up/in
9.77	97	0	like/vb	–7.17	0	52	related/vbn
9.67	103	2	enough/qlp	–7.10	0	51	due/jj
9.46	156	20	not/*	–6.96	0	49	attention/nn
9.40	90	0	likely/jj	–6.60	31	95	came/vbd
9.14	93	2	tried/vbd	–6.28	0	40	regard/nn
8.95	107	7	seem/vb	–6.28	0	40	approach/nn
8.80	83	1	expected/vbn	–6.20	0	39	relation/nn
8.51	74	0	try/vb	–6.03	0	37	next/in
8.09	67	0	ready/jj	–5.78	0	34	return/vb
8.08	85	5	as/cs	–5.77	1	36	lead/vb
8.05	74	2	difficult/jj	–5.69	0	33	prior/rb
8.03	66	0	how/wrb	–5.69	3	39	said/vbd

Thus, we see there is considerable leverage to be gained by preprocessing the corpus and manipulating the inventory of tokens. Unfortunately, the Tagged Brown Corpus is fairly small (N = 1 million words). If one wanted a very long list of verbs that take one use of *to* more than the other, one would need a much larger corpus. We have used the automatic tagger described in (Church, 1988) to tag the 1988 AP corpus and to meet this need.

Before moving onto the text topic, preprocessing with a parser, let's consider one more example: the contrast between the subordinate conjunction *that* and the demonstrative pronoun *that*. Suppose that one wanted to construct a set of rules for disambiguating the two. Then it might be helpful to look at the words on either side and see which of them gives the most leverage. Again, it is helpful to work from a tagged corpus such as the Tagged Brown Corpus.

These tables are also useful for identifying errors in the tagged corpus. When there are just a few exceptions to an overwhelming pattern, there is a good chance that the exceptions are really mistakes. Consider *so that*. In all but two cases, the *that* is a subordinate conjunction. The concordances (below) show quite clearly that the two so-called exceptions are really errors in the Tagged Brown Corpus. Similarly, the exceptions to the *fact that* pattern are also probably mistakes.

Table 6: Which words follow the subordinate conjunction (that/cs) and which words follow the demonstrative pronoun (that/dt)?

subordinate conjunction				demonstrative pronoun			
t	that/cs w	that/dt w	w	t	that/cs w	that/dt w	w
36.30	1346	2	the/at	–12.50	3	159	of/in
21.44	529	6	he/pps	–10.49	0	110	is/bez
17.42	320	1	it/pps	–9.28	0	86	./.
15.93	259	0	they/ppss	–7.95	2	65	time/nn
13.85	197	0	a/at	–7.81	0	61	was/bedz
11.61	140	0	she/pps	–7.64	3	61	’’/’’
11.57	139	0	I/ppss	–6.25	0	39	way/nn
11.30	133	0	this/dt	–5.83	0	34	day/nn
11.10	139	1	we/ppss	–5.30	0	28	which/wdt
10.93	145	2	there/ex	–4.59	0	21	year/nn
8.71	81	0	his/pp$	–4.36	0	19	night/nn
7.82	76	1	all/abn	–3.92	92	72	,/,
7.50	71	1	if/cs	–3.61	0	13	moment/nn
7.14	56	0	these/dts	–3.49	1	13	matter/nn
6.52	98	7	‘‘/‘‘	–3.47	0	12	kind/nn
6.09	42	0	an/at	–3.47	0	12	would/md
6.00	50	1	no/at	–3.32	0	11	morning/nn
5.93	40	0	their/pp$	–3.17	0	10	?/.
5.82	113	12	in/in	–3.03	1	10	point/nn

Which words precede the subordinate conjunction (that/cs) and which words precede the demonstrative pronoun (that/dt)?

subordinate conjunction				demonstrative pronoun			
t	w that/cs	w that/dt	w	t	w that/cs	w that/dt	w
14.19	227	2	so/cs	–12.25	1	151	of/in
11.33	179	5	fact/nn	–9.31	17	102	in/in
7.47	86	3	say/vb	–9.00	0	81	to/in
7.27	67	1	believe/vb	–7.88	0	62	like/cs
6.73	50	0	clear/jj	–6.63	0	44	for/in
5.96	70	4	said/vbd	–6.48	0	42	at/in
5.77	38	0	realize/vb	–5.83	0	34	with/in
5.51	35	0	think/vb	–5.69	4	36	than/cs
5.42	34	0	evidence/nn	–5.57	0	31	from/in
5.42	34	0	felt/vbd	–5.20	0	27	on/in
5.39	56	3	knew/vbd	–5.20	0	27	do/do
5.14	31	0	indicate/vb	–5.20	0	27	about/in
5.04	30	0	found/vbd	–4.81	2	25	as/cs
4.95	29	0	now/rb	–4.25	1	19	But/cc
4.85	28	0	assume/vb	–4.00	0	16	by/in
4.74	48	3	show/vb	–4.00	0	16	At/in
4.42	24	0	says/vbz	–3.94	4	19	all/abn
4.31	23	0	means/vbz	–3.67	30	35	--/--
4.20	22	0	indicated/vbd	–3.50	2	14	after/in
4.20	22	0	true/jj	–3.47	0	12	that/cs

so	at annual/jj Christmas/np bazaar/nn ,/, so/cs	that/dt dusk/nn was/bedz beginning/vbg to/to gather/vb
that	and/cc down/rp from/in the/at table/nn so/cs	that/dt talk/nn was/bedz impossible/jj ./. Well/uh ,/,
fact	and/cc further/rbr embossed/vbd the/at fact/nn	that/dt baseball/nn rightfully/rb is/bez the/at nati
that	it/pps will/md probably/rb be/be the/at fact/nn	that/dt SEATO/nn forces/nns are/ber ready/jj to/to a
	n/jj beings/nns arise/vb from/in the/at fact/nn	that/dt man/nn is/bez not/* one/cd ,/, but/cc many/a
	tent/jj ;/. for/cs ,/, using/vbg the/at fact/nn	that/dt .ul/1000 N/nn .ul/0 and/cc .ul/1000 N'/nn .u
	/nn-tl ./. .PP/ Aside/rb from/in the/at fact/nn	that/dt business/nn was/bedz slow/jj this/dt time/nn

Thus, we have seen three applications where one would be interested in preprocessing the corpus with a part of speech tagger in order to highlight the distribution of parts of speech. First, the tagger and t-score combination can be used to help the grammar writer design disambiguation rules. Secondly, the combination can be used to improve the coverage of complement structures. And thirdly, the combination can be used as a sanity check to spot likely sources of errors in the tagged corpus. We now turn our attention to preprocessing the corpus with a parser in order to highlight relationships between subjects, verbs and objects.

3.2 On the Interaction between Syntax, Semantics and Statistics

Chomsky argues quite convincingly that syntax should play an important role in the interpretation of semantic and statistical factors.

> ''We return to the question of the relation between semantics and syntax in sections 8,9, where we argue that this relation can only be studied after the syntactic structure has been determined on independent grounds. I think that much the same thing is true of the relation between syntactic and statistical studies of language. Given the grammar of a language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the use of language (as distinct from the syntactic structure of language) can be quite rewarding.... [Chomsky then cites articles by Mandelbrot and Simon, who were debating the statistics behind Zipf's Law.]
>
> One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable, but I know of no suggestion to this effect that does not have obvious flaws.'' (Chomsky 1957, p. 17, footnote 4)

Chomsky's suggestion that statistical preferences might be applied after syntactic analysis is an extremely intriguing one. We would prefer to rephrase his suggestion slightly, though. It really isn't necessary to first parse and then interpret in order to capture the spirit of his suggestion. The order of application is an implementational detail. As a practical matter, once the preferences are thrown into a chart parser (dynamic program), it is sometimes very hard to tell what really happens before what. The crucial point is that the statistics should depend on the syntactic context.

Of course, it probably isn't practical to do this ''right,'' given limited computing resources and available training material, and therefore, we will have to adopt some simplifying assumptions (that have some obvious flaws). Nevertheless, we believe that the tools that will be proposed here can produce usable results which at the very least provide a starting point for lexical analysis.

The idea of a stochastic context-free grammar is not new; there were at least three such suggestions at the International Workshop on Parsing Technologies, held in 1989 at CMU: Fujisaki et al. (1989), Seneff (1989) and Su et al. (1989). There are also many older references such as Suppes (1970). Most of these

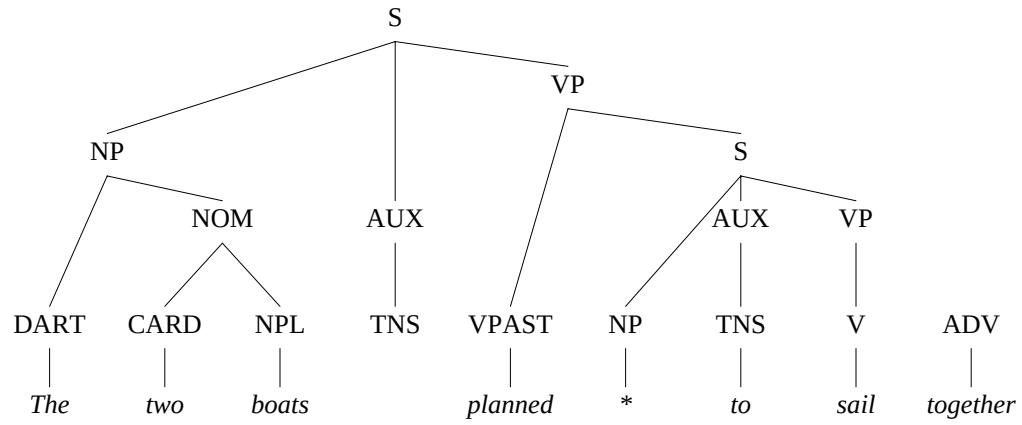
proposals, though, do not adequately model preferences among lexical items, in our opinion. It is very difficult, in these systems, to take advantage of the fact that the subject *boat* is likely to be found with the verb *sail*.

3.3 Preprocessing the Corpus with a Parser

We have decided to focus on lexical preferences. In particular, let us consider the question, ‘‘What does a boat typically do?’’ More specifically, we want to make a list of the verbs that are most associated with the subject *boat*. The first step in the process is to parse the corpus using a syntactic parser (Fidditch) designed to provide surface analyses of unrestricted text. Given the current state of the art, the parser makes many errors, but despite the error rate, many interesting distributional relations emerge.

Look in a little more detail at what the parser does with the sentence fragment:

(1) The two boats planned to sail together ...



The parser managed to reconstruct the surface structure for this sentence fragment without problems, although the adverb *together* is left unattached. The parser identified both the main clause, [S *The two boats planned S*], and the subordinate clause, [S *to sail*]. It found that the main clause contains the subject, [NP *The two boats*], and the verb, [VPAST *planned*]. The parser also identified the subordinate clause as the complement of the verb *plan*, by making use of the fact that the lexical entry for *plan* indicates that such an infinitival complement is possible. The parser’s lexicon contains a description of the complement structures for *plan* and about 700 other common verbs.

This parse tree is then reduced to two SVO triples:

1. boat/S plan/V ?/O
2. boat/S sail/V ?/O

In order to propose these SVO triples, the parser had to determine that the subject of the main clause, [S *The two boats planned S*], is also the subject of the subordinate clause, [S *to sail together*]. In addition, the parser identified *boats* as the head of the noun phrase, and *boat* as the base form of *boats*. Similarly, *plan* was identified as the base form of the head of the verb phrase in the main clause, and *sail* was identified as the base form of the head of the verb phrase in the subordinate clause. In this example, no direct object was found for either clause (which is fortunate, since there shouldn’t be one). In more difficult examples, it is necessary to undo the effects of various transformations including passive and wh-movement.

Let's consider some more difficult examples where the parser makes some mistakes. Table 7 shows 29 SVO triples that were extracted from the AP story, *Three Drowned, 1 Missing in Boat Accident*. As you can see, the parser manages to find the majority of SVO triples, though there are plenty of errors. The mistake of identifying *today* as the object of *search* derives from the lookup table of parts of speech, where *today*, along with *tomorrow* and *yesterday*, is classified as a noun. British learners' dictionaries classify these words as adverbs.

A typical parsing mistake is that *boat* is not in fact the subject of *drown*. However, it is easy to see how the parser made this mistake: it was confused by the dangling participial phrase *drowning his baby sister*. Thus, the triples are by no means perfect. There are other interesting parsing errors, too. Nevertheless, we have found that in practice the triples yielded by this parser are good enough to provide considerable information about the lexical preferences among subjects, verbs and objects.

The 44 million word 1988 AP corpus was parsed (in about 16 days of computer time on a nicely loaded Sun4), producing a set of 8,225,886 subject-verb-object (SVO) triples. From these 8,225,885 million triples, we constructed three times as many SV, SO and VO pairs.[9] Thus, N = 24,677,658. We also computed $P(x, y)$, $P(x)$ and $P(y)$, by simply counting the frequencies $f(x, y)$, $f(x)$ and $f(y)$ in the set of N pairs, and dividing by N. We apply the mutual information statistic to these 24.7 million pairs in order to identify interesting associations among subjects, verbs and objects, as shown in table 8 (below).

Table 8 lists all verbs that the parser found at least three times in this 44 million word corpus with the subject *boat*. (The arbitrary cut-off of three is introduced here in order to alleviate the fact that mutual information values are misleading when the frequency counts are small, unless special care is taken.)

9. This step actually adds $\log_2 0.75 \approx 0.42$ to the mutual information values. To compensate, we should subtract 0.42 bits from the values in table 8.

(2) The Coast Guard searched today for a 5-year-old boy missing from an overloaded boat that capsized in New Bedford Harbor, drowning his baby sister, their mother and another woman...

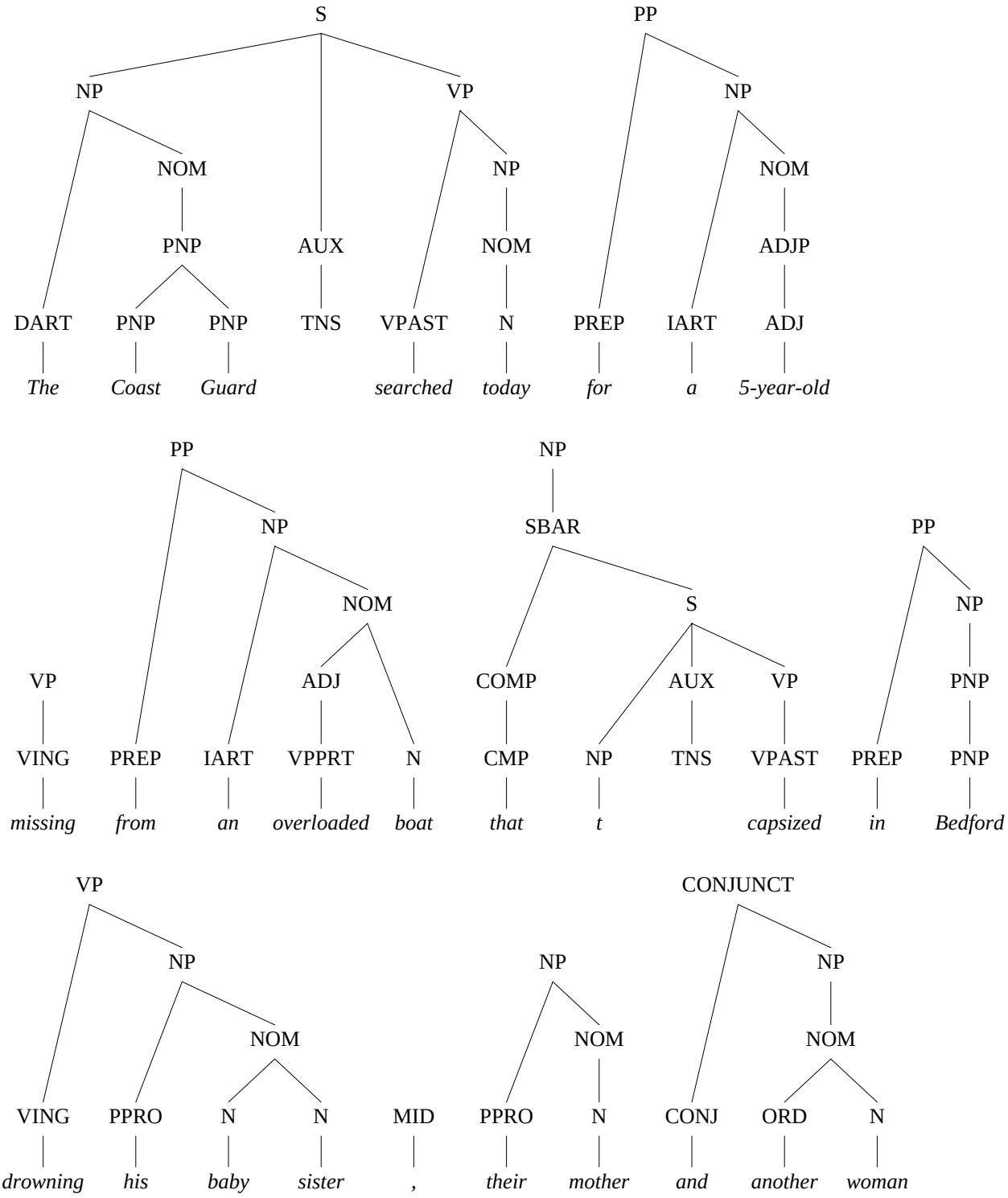


Table 7: SVO Triples in *Three Drowned, 1 Missing in Boat Accident* (July 5, 1988)

SVO Triple	Text
Guard/S search/V today/O	NEW BEDFORD, Mass (AP) -- The Coast Guard searched today for a 5-year-old boy
boat/S capsize/V ?/O	missing from an overloaded boat that capsized in New Bedford Harbor,
boat/S drown/V sister/O	drowning his baby sister, their mother and another woman,
official/S say/V ?/O	officials said.
PASSIVE/S throw/V people/O	Fifteen people were thrown into the water Monday night
boat/S return_from/V Fourth/O	as the 22-foot boat was returning from a Fourth of July fireworks display,
Monday/S say/V Foley/O	said Coast Guard Petty Officer David Foley.
survivor/S say/V boat/O	Survivors said
?/S capsize/V ?/O	the boat capsized
it/S hit/V wake/O	when it apparently hit another boat's wake
Foley/S say/V ?/O	while making its way through heavy fog, Foley said.
PASSIVE/S pull/V people/O	Eleven people were pulled safely from the water,
he/S say/V ?/O	he said.
PASSIVE/S overload/V boat/O	The boat ''was definitely overloaded,''
?/S say/V Foley/O	said Foley.
?/S rescue/V survivor/O	Coast Guard vessels and private craft rescued the survivors.
fog/S be/V problem/O	''The fog was the main problem,''
Foley/S say/V ?/O	Foley said.
There/S be/V visibility/O	''There was no visibility.
PASSIVE/S sock/V it/O	It was really socked in.''
Divers/S find/V body/O	Divers found the body of a 1-year-old girl under the bow of the boat and the bodies of her mother and another woman under the stern.
?/S identify/V victim/O	Louise Hathaway, night supervisor at St. Luke's Hospital in New Bedford, identified the victims as Jean Fauteaux, 50; Maria Carvalho, 27; and her daughter, Melissa, 15 months, all of North Dartmouth.
she/S say/V brother/O	She said Melissa's 5-year-old brother, Kenneth Carvalho,
?/S miss/V ?/O	was missing.
PASSIVE/S admit/V ?/O	Their sister, Amy Lynn Carvalho, 8, and Allan Viera Jr., 1, of New Bedford were admitted to the hospital.
Hathaway/S say/V ?/O	Asked their conditions, Hathaway said,
They/S do/V ?/O	''They'll do fine.''
PASSIVE/S treat/V people/O	Seven people were treated at the hospital and released,
she/S say/V ?/O	she said.

Table 8: What does a boat do?
(N = 24,677,658; f(x, y) ≥ 3).

I(x;y)	f(x,y)	f(x)	f(y)	x	y	I(x;y)	f(x,y)	f(x)	f(y)	x	y
11.01	16	984	194	boat/S	capsize/V	3.09	4	984	11768	boat/S	fail/V
9.30	51	984	2036	boat/S	sink/V	2.72	4	984	15244	boat/S	stop/V
8.17	3	984	262	boat/S	cruise/V	2.59	5	984	20894	boat/S	accord/V
7.40	6	984	890	boat/S	sail/V	2.54	4	984	17266	boat/S	reach/V
7.27	3	984	488	boat/S	tow/V	2.14	3	984	17074	boat/S	lose/V
7.18	3	984	518	boat/S	turn_in/V	2.09	6	984	35456	boat/S	leave/V
6.83	3	984	660	boat/S	collide/V	2.04	4	984	24410	boat/S	keep/V
6.61	3	984	772	boat/S	drown/V	2.04	6	984	36494	boat/S	kill/V
6.34	4	984	1238	boat/S	drag/V	1.69	6	984	46624	boat/S	be_in/V
6.28	3	984	968	boat/S	escort/V	1.61	3	984	24714	boat/S	put/V
6.04	4	984	1522	boat/S	overturn/V	1.38	8	984	77238	boat/S	take/V
5.90	5	984	2096	boat/S	rescue/V	1.36	3	984	29338	boat/S	hold/V
5.43	5	984	2902	boat/S	approach/V	1.28	4	984	41232	boat/S	use/V
4.64	16	984	16068	boat/S	carry/V	1.26	3	984	31506	boat/S	become/V
4.43	9	984	10470	boat/S	hit/V	0.94	19	984	247542	boat/S	have/V
4.18	4	984	5524	boat/S	travel/V	0.67	3	984	47214	boat/S	begin/V
3.86	6	984	10348	boat/S	pass/V	0.57	3	984	50766	boat/S	get/V
3.71	4	984	7656	boat/S	attack/V	0.17	4	984	89256	boat/S	do/V
3.48	3	984	6748	boat/S	injure/V	–0.35	26	984	830120	boat/S	be/V
3.38	4	984	9614	boat/S	fire/V	–0.35	3	984	95880	boat/S	make/V
3.30	3	984	7634	boat/S	operate/V	–3.38	4	984	1045494	boat/S	say/V

First, we note that mutual information ranks the verbs as we intuitively expect: it shows that *boat* is an interesting subject for the verb *sail* but not for the verb *be*, because the mutual information values are 7.40 and – 0.35 respectively. This ranking accords with our intuitions that *boat* is associated more with *sail* than with *be*. Note that this association is not revealed by an alternative measure, namely the raw frequency: *boat* occurs 26 times as subject of *be*, but only 6 times as subject of *sail*. We further observe that the verbs at the top of the list, those with relatively high mutual information, tend to characterize what boats typically do.

A lexicographer should now scan the verbs at the top of the list and check for verbs such as *drown* that seem intuitively implausible: typically, *boats* do not *drown* (intransitive), nor do they (transitively) *drown* people. Since there are only a few cases like *drown* where the mutual information table is misleading, it shouldn't be too much trouble for the lexicographer to check back to the original text and determine whether what is at issue is a parsing error, a piece of loose prose, or an unusual use of a familiar word.

A table of SVO associations such as this could have a number of important uses. First, we would hope that we could use the associations in order improve future parsers. Hopefully, a parser could someday use the SVO associations in order to predict that *boat* is probably not the subject of *drowning* in sentence (2).

Secondly, we would hope that we could present these associations to lexicographers in a way so that they would have an easier time partitioning concordance lines into senses. In addition, the SVO associations might prove helpful in budgeting resources for the concordance analysis phase. Note that the pairs near the top of the SVO association involve words that lexicographers consider to be fairly easy, in contrast with the words toward the bottom of the list. *boat, sail, capsize* and *cruise* are considered to be fairly easy words because they don't have very many different dictionary senses. Some of the difference is due to the fact

that the words farther down the list are more frequent, and frequent words are usually more complex and therefore harder to analyze accurately.

Before moving on to the next topic (step 3: selecting the appropriate unit of text), let's consider one more example. What do you typically do with *food* and *water*? This example will illustrate the use of the mutual information and t-score statistics on SVO triples.

Notice how the evidence in table 9 (below) draws attention to some of the cultural facts about *food* and *water* in American English, facts, which are socially salient, though not perhaps psychologically salient (ie., not immediately obvious via introspection). *Food*, for example, is a valuable commodity, which people *hoard*, *donate* or *buy* as the case may be. The question of *hoarding*, *donating* or *buying water*, on the other hand, (in American culture, at least) does not often arise. *Water* is more typically *polluted*, *contaminated* and *poisoned* in AP news stories.

Table 9: What do you typically do with *food* and *water*?
Computed over Parsed AP Corpus (N = 24.7 million SVO triples)

Associated with food						Associated with water					
I(x;y)	fxy	fx	fy	x	y	I(x;y)	fxy	fx	fy	x	y
9.62	6	84	2240	hoard/V	food/O	9.05	16	208	3574	conserve/V	water/O
8.83	9	218	2240	go_without/V	food/O	8.98	18	246	3574	boil/V	water/O
7.68	58	3114	2240	eat/V	food/O	8.64	6	104	3574	ration/V	water/O
6.93	8	722	2240	consume/V	food/O	8.45	10	198	3574	pollute/V	water/O
6.42	6	772	2240	run_of/V	food/O	8.40	20	408	3574	contaminate/V	water/O
6.29	14	1972	2240	donate/V	food/O	8.37	38	794	3574	pump/V	water/O
6.08	17	2776	2240	distribute/V	food/O	7.86	6	178	3574	walk_on/V	water/O
5.14	51	15900	2240	buy/V	food/O	7.81	43	1320	3574	drink/V	water/O
4.80	53	21024	2240	provide/V	food/O	7.39	15	618	3574	spray/V	water/O
4.65	13	5690	2240	deliver/V	food/O	7.39	9	370	3574	poison/V	water/O

If we wanted to contrast the verbs that take *food* with those that take *water*, it might be helpful to use the t-score argument that we used for contrasting the words after *strong* and *powerful*. These are given in table 10. Note that the verbs with extreme t-scores seem very natural. As suggested earlier, we would hope that a parser could someday make use of facts such as these. A parser really ought to be able to take advantage of the fact that *eating food* and *drinking water* are much more plausible that *eating water* and *drinking food*, but without a tool such as the t-score, it is just too labor-intensive to deal with facts such as these.

However, to a lexicographer, the t-scores for *food* and *water* are not very interesting, because the words are semantically very far apart. Dictionary users, unlike dumb computers, have enough common sense to know that *food* is typically *eaten* and *water* is typically *drunk*. To a lexicographer, the contrast between *strong* and *powerful* is much more interesting than the contrast between *food* and *water*, because *strong* and *powerful* are so close in meaning that it isn't obvious how they differ. In order to use the t-test effectively, the lexicographer needs to pick a pair of words like *strong* and *powerful*; if you pick a pair like *food* and *water*, then the t-score won't tell you anything you didn't already know.

Table 10: What do you do more with *food* than with *water*?
Computed over Parsed AP Corpus (N = 24.7 million SVO triples)

More with food				More with water			
t	food	water	w	t	food	water	w
7.47	58	1	eat/V	–6.93	0	50	be_under/V
6.26	51	7	buy/V	–5.62	1	38	pump/V
4.61	31	6	include/V	–5.37	3	43	drink/V
4.47	53	25	provide/V	–5.20	0	29	enter/V
4.18	31	9	bring/V	–4.87	1	30	divert/V
3.98	21	3	receive/V	–4.80	0	25	pour/V
3.69	14	0	donate/V	–4.25	0	20	draw/V
3.55	13	0	prepare/V	–4.01	0	18	boil/V
3.31	13	1	offer/V	–3.89	0	17	fall_into/V
3.08	13	2	deliver/V	–3.75	1	20	contaminate/V

This table reinforces the observation we made above that the t-score and the mutual information score are addressing very different questions. Note that verbs with extreme t-scores are generally quite frequent. In contrast, verbs with extreme mutual information scores are generally quite infrequent. Note, for example, the difference between *eat* and *hoard*.

4. Step 3: Select Appropriate Unit of Text

4.1 Discourse Context

In the previous section, we showed how the statistical tools could be used with a parser in order to discover interesting relationships between predicates and arguments, an example of step 2 (preprocessing). In this section, we begin the discussion of step 3 (selecting an appropriate unit of text) by showing how statistics computed on discourses differ from those computed over bigrams and clauses. Let us now consider the information retrieval application, where it is desirable to compute the statistics over discourse units.

Table 11 (below) uses the t-score to contrast words that appear in the same AP story as the word *food* with words that appear in the same AP story as the word *water*. As you can see, the words that are found in *food* stories do seem to be ‘‘associated’’ more with *food* than with *water*, but the associations are different from the ones in the previous section. Now, the statistics are latching onto newsworthy topics such as food prices and accidents at sea.

The second column in table 11 gives $f(\mathit{food}, w)$, the number of stories in the 1988 AP corpus that mention *food* and *w*. The third column in table 11 gives $f(\mathit{water}, w)$, the number of stories that mention *water* and *w*. The first column gives the t-score:

$$t \equiv \frac{P(w|\mathit{food}) - P(w|\mathit{water})}{\sqrt{\sigma^2(w|\mathit{food}) + \sigma^2(w|\mathit{water})}}$$

where the probabilities and variances are computed with the ELE. That is,

$$P(w|food) \approx \frac{f(w\ food) + 1/2}{f(food) + V/2}$$

$$P(w|water) \approx \frac{f(w\ water) + 1/2}{f(water) + V/2}$$

$$\sigma^2(w|food) \approx \frac{f(w\ food) + 1/2}{(f(food) + V/2)^2}$$

$$\sigma^2(w|water) \approx \frac{f(w\ water) + 1/2}{(f(water) + V/2)^2}$$

V is the number of words that appear with either *food* or *water*. In this case, $V = 111{,}635$, $f(food) = 1.31$ million and $f(water) = 1.24$ million.

Table 11: How do stories that mention *food* differ from stories that mention *water*?

More like food				More like water			
t	food	water	w	t	food	water	w
50.74	4174	611	food	–51.11	611	4052	water
16.90	491	80	consumer	–11.31	105	335	crew
16.30	610	149	products	–11.47	91	316	inches
15.90	740	228	prices	–11.58	113	356	environmental
14.67	423	85	goods	–11.73	99	337	river
14.28	383	72	Food	–12.12	42	244	pollution
13.91	402	87	stock	–12.16	41	243	Water
13.85	665	233	market	–12.23	578	1034	near
13.41	341	65	inflation	–12.40	183	493	rain
12.99	359	80	clothing	–14.35	649	1231	miles
12.94	585	206	price	–15.34	189	609	River
12.88	264	37	takeover	–16.56	238	739	feet
12.44	489	160	sales	–11.28	80	292	Lake
12.39	482	157	rose	–10.99	316	635	air
12.29	781	345	economic	–10.94	119	347	Coast
12.09	290	59	consumers	–10.93	78	279	Navy
12.06	184	13	earnings	–10.90	43	215	gallons
11.90	305	69	trading	–10.76	37	200	vessel
11.89	462	155	share	–10.69	101	311	boat
11.74	676	291	increase	–10.66	107	320	waters
11.72	490	175	economy	–10.53	66	248	accident
11.67	161	8	buyout	–10.51	128	349	sea
11.63	244	43	shares	–10.24	115	321	coast
11.56	334	89	rebels	–10.16	98	292	ship

This sort of t-score tool might be a useful adjunct to a keyword information retrieval system. In a keyword system, it is often hard to deal easily with a working set of a few hundred or a few thousand documents. It might be very useful to provide the user with tools that would generate a good set of candidate keywords that he might want to try. Suppose, for example, that the user had selected the keyword *food*, and discovered that there were 4174 *food* stories, which is much more than he wanted. He might then try to

narrow down the set by picking a word like *water*, which is likely to pick out a sense of *food* that he doesn't want. Then after constructing a table like table 11 above, he could see that he should refine his query to select for stories that mention some of the words with large t-scores and to reject stories that mention some of the words with very small t-scores. Of course, he will need to use some common sense in editing the list down to something more reasonable. The list won't be perfect, but it should be good enough that the editing job isn't too time consuming. It is probably easier to edit down a list that is slightly too long than to start from scratch.

It might also be interesting to a lexicographer to use the t-score tool on discourse size units of text. But the lexicographer would probably want to look at a different pair of words, where the contrast is more subtle than between *food* and *water*. Consider, for example, the pair *boat* and *ship*. The Cobuild dictionary (Sinclair et al., 1987) gives a fairly good sense of the difference:

- *boat*: a small vessel for travelling on water, especially one which only carries a few people.
- *ship*: a large boat which carries passengers or cargo on sea journeys.

The basic difference, that a ship is bigger than a boat, accounts for many of the t-scores in table 12 (below). (In order to save space, only a small fraction of the significant words are shown in the table; there are almost 4000 words that are significant at the 95% confidence level.) Nevertheless, the small sample that is shown in the table below gives a fairly good sense of the difference between *boat* and *ship*. A *boat* is generally smaller than a *ship* in some sense, but there are quite a number of different senses that might apply. For example, the table shows that boats are found on rivers and lakes, whereas ships are found in the Mediterranean Sea and near Iran. Boats are also used for small jobs (e.g., fishing, police, pleasure), whereas ships are used for serious business (e.g., hauling valuable cargo and fighting wars). People are also more likely to drown in boats than on ships, as evidenced by the references to the Vietnamese boat people.

4.2 Polysemy

Polysemy is a well-known problem for keyword systems. Suppose that a user wanted to find stories that mention *bank*, but only in the ''money'' sense and not in the ''river'' sense. We think it would be useful to design an interactive tool that could help the user focus in on one sense or the other.

Table 13 (below) shows that the t-test could be of some use. The key here is step 3, selecting the appropriate unit of text. The table was computed from a set of 45 stories that mentioned both *bank* and *river*, and another set of 467 stories that mentioned both *bank* and *money*. In this case, $V = 21,585$, $f(\textit{bank \& river}) = 12,923$ and $f(\textit{bank \& water}) = 136,231$.

Note how well the examples in table 13 agree with the examples cited by the third author.

> ''On the one hand, *bank* co-occurs with words and expressions such as *money, notes, loan, account, investment, clerk, official, manager, robbery, vaults, working in a, its actions, First National, of England*, and so forth. On the other hand, we find *bank* co-occurring with *river, swim, boat, east* (and of course *West* and *South*, which have acquired special meanings of their own), *on top of the*, and *of the Rhine*.'' (Hanks 1987, p. 127)

Table 12: The Difference between Boat and Ship in 1988 AP Newswire

Boat				Ship			
t	boat	ship	w	t	boat	ship	w
30.2	1999	251	ship	–29.5	251	1210	boat
12.5	310	34	USS	–11.0	36	171	Vietnamese
10.7	502	115	Navy	–10.4	48	170	refugees
10.6	282	41	sailors	–9.8	193	302	boats
9.7	145	10	Pentagon	–9.7	168	278	fishing
9.4	163	16	carrier	–9.2	21	114	Kong
9.3	386	89	WASHINGTON	–9.0	24	114	Hong
9.3	103	3	turret	–8.7	11	91	persecution
9.1	124	8	battleship	–8.2	7	78	repatriation
8.9	328	72	tanker	–8.1	34	110	refugee
8.7	446	119	ships	–7.7	5	66	HONG
8.5	155	19	Iowa	–7.7	5	66	KONG
8.4	222	40	explosion	–7.7	75	146	Vietnam
8.3	267	56	gallons	–7.5	796	705	people
7.9	222	44	aground	–7.3	23	84	camps
7.8	253	56	aircraft	–7.0	25	81	colony
7.5	186	35	crude	–6.9	32	88	drowned
7.4	141	21	Adm.	–6.8	8	58	Refugees
7.4	281	70	spill	–6.7	35	88	homeland
7.3	125	17	guns	–6.7	91	142	fishermen
7.2	106	12	Fleet	–6.6	40	91	river
7.2	267	67	cargo	–6.5	58	108	fled
7.0	74	5	Iranian	–6.4	72	120	woman
6.9	69	4	16–inch	-6.3	10	54	fisherman
6.8	92	10	shipments	–6.3	208	233	police
6.7	51	1	Hartwig	–6.2	36	81	High
6.7	160	32	Cmdr.	–6.1	5	45	resettlement
6.7	207	49	Exxon	–5.8	16	54	immigrants
6.7	433	142	oil	–5.7	15	52	detention
6.6	103	14	blast	–5.7	21	59	fleeing
6.6	349	107	military	–5.7	28	66	prove
6.6	95	12	Norfolk	–5.6	125	152	missing
6.6	49	1	gunner	–5.6	3	36	Hanoi
6.5	91	11	Iran	–5.5	19	54	asylum
6.5	121	20	gun	–5.5	446	394	water
6.5	123	21	missile	–5.4	41	75	Lake
6.4	128	23	Mediterranean	–5.3	140	160	River
6.4	318	96	Soviet	–5.2	155	170	rescued
6.4	99	14	Hazelwood	–5.2	427	371	back
6.4	95	13	bomb	–5.1	24	55	immigration

Table 13: Sense Disambiguation

River Sense of Bank				Money Sense of Bank			
t	bank & river	bank & money	w	t	bank & river	bank & money	w
6.63	45	4	river	–15.95	6	467	money
4.90	28	13	River	–10.70	2	199	Bank
4.01	20	13	water	–10.60	0	134	funds
3.57	16	11	feet	–10.46	0	131	billion
3.46	23	39	miles	–10.13	0	124	WASHINGTON
3.44	21	32	near	–10.13	0	124	Federal
3.27	12	5	boat	–9.43	0	110	cash
3.06	14	16	south	–9.03	1	134	interest
2.83	8	1	fisherman	–8.79	1	129	financial
2.83	21	49	along	–8.79	0	98	Corp
2.76	11	12	border	–8.38	1	121	loans
2.74	17	35	area	–8.17	0	87	loan
2.72	9	6	village	–7.57	0	77	amount
2.71	7	0	drinking	–7.44	0	75	fund
2.70	16	32	across	–7.38	0	74	William
2.66	9	7	east	–7.36	1	102	company
2.58	7	2	century	–7.31	1	101	account
2.53	10	13	missing	–7.25	0	72	deposits
2.52	6	0	Perez	–7.25	0	72	assets
2.52	6	0	barges	–7.12	0	70	raised
2.50	9	10	southern	–7.12	0	70	savings
2.49	13	25	saw	–7.12	0	70	attorney
2.45	6	1	dig	–7.08	1	97	paid
2.45	7	4	troops	–7.05	0	69	House
2.41	8	8	covered	–7.02	1	96	business
2.41	8	8	population	–6.92	0	67	prices
2.38	6	2	commander	–6.85	0	66	investigation
2.38	6	2	yards	–6.83	3	133	pay
2.37	10	16	fire	–6.82	4	150	percent
2.33	7	6	port	–6.79	1	92	asked
2.33	7	6	soldiers	–6.78	0	65	YORK
2.31	5	0	artifacts	–6.71	0	64	firm
2.31	5	0	riverbank	–6.64	0	63	Attorney
2.31	5	0	villagers	–6.50	0	61	investment
2.31	6	3	fish	–6.43	0	60	stock
2.31	13	29	military	–6.43	0	60	industry
2.24	6	4	bottom	–6.43	0	60	estate
2.23	5	1	ancient	–6.43	0	60	debt
2.23	5	1	swim	–6.43	0	60	agreed
2.22	9	15	dead	–6.41	4	141	federal

4.3 An Aside: Upper Case vs. Lower Case

It is also interesting to point out that the distinction between upper and lower case is sometimes very useful. Note that *Bank*, in table 13 above, is very strongly associated with the money sense and not with the river sense. It turns out that if *bank* and *Bank* appear in the same AP story, it is extremely likely that both of them will refer to a money bank (and not to some other sense such as *the West Bank*).

It is also interesting that the results are quite different if we compare stories that mention *Bank* with those that mention *bank*. In table 14 (below), we see a strong contrast between *Bank* (as in *the West Bank*) and the more common money sense of *bank*. The reason for the difference is that *the West Bank* stories do not tend to mention the word *bank*, spelled with a lower case *b*. Thus, we see that it makes sense to maintain the distinction between upper and lower case, at least for some applications. We consider the issues of whether or not to collapse upper and lower case to be a subcase of the general issue of preprocessing the corpus (step 2).

Table 14: Upper Case vs. Lower Case

Bank				bank			
t	Bank	bank	w	t	Bank	bank	w
35.02	1324	24	Gaza	–36.48	1284	3362	bank
34.03	1301	36	Palestinian	–10.93	900	1161	money
33.60	1316	48	Israeli	–10.43	624	859	federal
33.18	1206	26	Strip	–9.59	586	786	company
32.98	1204	29	Palestinians	–8.47	282	430	accounts
32.68	1339	72	Israel	–8.26	544	693	central
31.56	4116	1284	Bank	–8.21	408	554	cash
31.13	1151	47	occupied	–8.12	675	816	business
30.79	1104	40	Arab	–7.74	546	676	loans
27.97	867	21	territories	–7.54	52	140	robbery

5. Conclusion

5.1 Summary

We have mentioned three steps requiring human judgment:

1. choose an appropriate statistic (e.g., mutual information, t-score),
2. preprocess the corpus to highlight properties of interest (with a part of speech tagger or a parser), and
3. select an appropriate unit of text (e.g., bigram, SVO triple, discourse).

First, it is important to choose the appropriate statistic for the application. We discussed mutual information and t-scores in some detail. Mutual information is better for highlighting similarity; t-scores are better for establishing differences among close synonyms. We wouldn't want to say that one statistic is better than the other; both are important. There are times when we are more interested in finding associations, and there are other times when we are more interested in focusing in on subtle distinctions.

Secondly, it is useful to preprocess the corpus (or transform the data) appropriately for the application. If one wants to study the distribution of predicates and their arguments, then it is extremely helpful to preprocess the corpus with a parser such as Fidditch. On the other hand, if one wants to look at the difference between the distribution of *Bank* and *bank* in AP stories, then such a transform would be ill-advised. We have discussed two methods of preprocessing the corpus: (1) tagging each word with a part of

speech, and (2) parsing each clause into an SVO triple.

Third, it is important to choose an appropriate unit of text. We have looked at bigrams, clauses and AP stories. The different size units yield different results, and answer different questions. Statistics on AP stories may be helpful to researchers interested in Information Retrieval; statistics computed on smaller units may be more helpful to syntacticians.

Finally, it is important to choose an appropriate sample. Although we haven't said very much about sampling, it is a major issue. Analyses of the AP corpus are likely to answer questions about the AP newswire; balanced corpora, such as the Brown Corpus, are probably more appropriate for answering questions about general language.[10] Like the other steps discussed above, it is probably best to select the sample to match the application (and the available resources).

5.2 Why Statistics?

We believe that for many purposes (for example, learners' dictionaries and natural language processing), it is desirable to focus on the ''central and typical'' facts of the language that every speaker is expected to know, and to stay clear of the gray area where the facts seem to be less clear cut. In the words of Bennett (1976, p. 5), ''perhaps we can bring order into the chaos by mastering some central and basic kinds of language-use, and then with their aid elucidating others.'' This approach has led to some successes, especially the Cobuild dictionary, which argued that dictionaries should try to describe use as well as meaning:

> ''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. No major uses are missed, and the number of times a use occurs has a strong influence on the way the entries are organized. Equally, the large group of texts, called the corpus, gives us reasonable grounds for omitting many uses and word-forms that do not occur in it. It is difficult for a conventional dictionary, in the absence of evidence, to decide what to leave out, and a lot of quite misleading information is thus preserved in the tradition of lexicography.'' (Sinclair et al, 1987, p. xv)

Our approach has much in common with a position that was popular in the 1950s. It was common practice to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Running through the whole Firthian tradition, for example, is the theme that ''You shall know a word by the company it keeps'' (Firth, 1957). Harris's ''distributional hypothesis'' dates from about the same period. He hypothesized that ''the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities'' (Harris

---

10. However, we have found that the AP corpus has two advantages that make it more suitable for answering certain questions about general language. First, it is much larger (and ever increasing). Secondly, it represents a fairly uniform source, so that it is easier to see what is conventional and to distinguish it from what is particular. A balanced corpus of only one million words, such as the Brown Corpus, represents so many different genres with such tiny samples that it is often impossible to see what the conventions of any particular genre may be. No doubt some of the language in the Brown Corpus is representative of the general conventions of English, some of particular genres, while some is peculiar to particular authors. But the samples are so small that it is often not possible to decide which is which. Sometimes it is more advantageous to have a large corpus from a single uniform source, rather than many small samples from many different sources.

1968, p. 12).

The interest in statistical approaches faded rather suddenly when Chomsky argued quite successfully that statistics should not play a role in his competence model.[11] Since he was interested in a set of questions that do not require preference judgments, he was probably justified in adopting the ‘‘competence approximation,’’ that all grammatical sentences be treated equally, and that preferences should be ignored. Obviously, this approximation simplifies matters greatly for the kinds of applications that he had in mind. It is probably true that the ‘‘competence approximation’’ is more appropriate for some applications and less appropriate for others.[12] The question is: how do we know if the competence approximation is appropriate for our application?

We fear that Chomsky’s argument may have been too effective in undermining the statistical approach, with insufficient attention to differences in research goals. For example, the emphasis on a sharp division between what ‘‘can’’ and ‘‘cannot’’ occur has been responsible for some confusion in many areas, including foreign language teaching. The notion that a language consists of a fixed set of rules and a finite set of lexical items, and that they can be *learned* by students, who will then ‘‘know’’ the language, is too attractive to be easily dislodged, but it has been responsible for some misconceptions *inter alia* about the status of infelicities typically produced by foreigners, which are sufficient to identify the speaker as foreign, although not actually wrong.

5.3 A Cautionary Note: Use the Right Tool for the Job

On the other hand, we believe that statistical approaches have to be used very carefully. We fear that statistics might become a fad, which could lead to considerable abuse. It is, of course, a natural tendency for a research community to become overly attached to a single approach and advocate it as the solution to all of the world’s problems, without seriously investigating the degree to which the assumptions underlying the approach are appropriate for a particular application. In this spirit, one might advocate a screwdriver as a universal tool and observe that it can be used to open cans. However, it is really inappropriate to abuse a screwdriver in this way (as evidenced by the large residuals that it leaves on the outside of the can); it is much more sensible to use the right tool for the job.

We think it is unlikely that any single tool could be appropriate for all problems in natural language. In this light, we would be suspicious of a proposal that advocated just one tool (such as a Hidden Markov Model (HMM) or a Neural Network) to solve all problems in natural language. Until we have a better understanding of when these tools are appropriate (and when they are not), it may be premature to attempt to use them in a self-organizing system. We believe that human judgment is required to select the

---

11. ‘‘[T]he notion ‘grammatical in English’ cannot be identified in any way with the notion ‘higher order of statistical approximation to English.’ It is fair to assume that neither sentence (1) [Colorless green ideas sleep furiously] nor (2) [Furiously sleep ideas green colorless] (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally ‘remote’ from English. Yet (1), though nonsensical, is grammatical, while (2) is not. Presented with these sentences, a speaker of English will read (1) with a normal sentence intonation, but he will read (2) ... with just the intonation pattern given to any sequence of unrelated words... Similarly, he will be able to recall (1) much more easily than (2), to learn it much more quickly, etc... Evidently, one’s ability to produce and recognize grammatical utterances is not based on notions of statistical approximations and the like. The custom of calling grammatical sentences those that ‘can occur,’ or those that are ‘possible,’ has been responsible for some confusion here... Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances. I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure.’’ (Chomsky 1957, pp. 15-17)

12. For example, in psycholinguistic studies, it is well-known that preferences are very important and cannot be ignored. There is a long literature establishing the effect of word frequencies and word association norms in predicting reaction times and error rates.

appropriate tool and make sure that it doesn't run amuck.

To end on an optimistic note, we believe that these tools could be put together into a very useful workbench that would dramatically enhance the productivity of lexicographers in the near term. We hope to automate away some of the drudgery of lexicography, and to make possible insights that would not have been possible otherwise, even given any amount of drudgery.

References

Box, G., and Tiao, G. (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts.

Chapman, R. (1977) *Roget's International Thesaurus*, fourth edition, Harper & Row, New York.

Chodorow, M., Byrd, R., and Heidorn, G. (1985) ''Extracting semantic hierarchies from a large on-line dictionary,'' ACL Proceedings.

Chomsky, N. (1956) ''Three Models for the Description of Language,'' IRE Transactions on Information Theory, vol. IT-2, Proceedings of the Symposium on Information Theory.

Chomsky, N. (1957) *Syntactic Structures*, The Hague: Mouton & Co.

Church, K. (1988) ''A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,'' Second Conference on Applied Natural Language Processing, Austin, Texas.

Church, K., and Hanks, P. (1989) ''Word Association Norms, Mutual Information, and Lexicography,'' ACL Proceedings. Also (to appear) in *Computational Linguistics*.

Cowie, A. (ed.) (1989) *Oxford Advanced Learner's Dictionary*, Oxford University Press.

DeRose, S. (1988) ''Grammatical Category Disambiguation by Statistical Optimization,'' *Computational Linguistics*, Vol. 14, No. 1.

Firth, J. (1957) ''A Synopsis of Linguistic Theory 1930-1955'' in *Studies in Linguistic Analysis*, Philological Society, Oxford; reprinted in Palmer, F. (ed. 1968) *Selected Papers of J.R. Firth*, Longman, Harlow.

Flexner, S., Hauck, L., et al. (eds.) (1987) *The Random House Dictionary of the English Language, Second Edition, Unabridged*, Random House, New York.

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

Fujisaki, T., Jelinek, F., Cocke, J., Black, E. (1989) *Probabilistic Parsing Method for Sentence Disambiguation*, presented at the International Workshop on Parsing Technologies, CMU.

Gale, W. and Church, K. (1990) ''What's Wrong with Adding One?'' submitted to *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

Good, I. J. (1953) ''The Population Frequencies of Species and the Estimation of Population Parameters,'' *Biometrika*, Vol. 40, pp. 237-264.

Gove, P., et al. (1961) *Merriam Webster's Third New International Dictionary*, Springfield, Mass., G. & C. Merriam Co.

Halliday, M. (1966) ''Lexis as a Linguistic Level,'' in Bazell, C., Catford, J., Halliday, M., and Robins, R. (eds.), *In Memory of J. R. Firth*, Longman, London.

Hanks, P. (1987) ''Definitions and Explanations,'' in Sinclair, J. (ed.) *Looking Up: An account of the COBUILD Project in lexical computing*, Collins, London and Glasgow.

Harris, Z. (1968) *Mathematical Structures of Language*, New York: Wiley.

Hindle, D. (1983) ''User manual for Fidditch, a deterministic parser,'' Naval Research Laboratory Technical Memorandum 7590-142

Jelinek, F. (1985) ''Self-organized Language Modeling for Speech Recognition,'' IBM Report.

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

Seaton, M., Davidson, G., Schwarz, C., Simpson, J. (1986) *Chambers 20th Century Thesaurus*, W &R Chambers Ltd, Edinburgh.

Seneff, S. (1989) *Probabilistic Parsing for Spoken Language Applications*, presented at the International Workshop on Parsing Technologies, CMU.

Sinclair, J., Hanks, P., Fox, G., Moon, R., Stock, P. et al. (eds.) (1987) *Collins Cobuild English Language Dictionary*, Collins, London and Glasgow.

Smadja, F. (1989) ''Macrocoding the Lexicon with Co-Occurrence Knowledge,'' in Zernik, U. (ed.) Proceedings of the First International Lexical Acquisition Workshop.

Su, K., Wang, J., Su, M. (1989) *A Sequential Truncation Parsing Algorithm Based on the Score Function*, presented at the International Workshop on Parsing Technologies, CMU.

Suppes, P. (1970) ''Probabilistic Grammars for Natural Languages,'' *Synthese* 22, pp. 95-116.