University of Brighton

ITRI-96-08

Which words are particularly characteristic of a text? A survey of statistical approaches

Adam Kilgarriff

April, 1996

Also published in *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex University, April 1996. Pp. 33–40

This work was supported by the EPSRC under Grant K18931, SEAL.

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4AT, UK
TEL: +44 1273 642900 EMAIL: firstname.lastname@itri.brighton.ac.uk
FAX: +44 1273 642908 NET: http://www.itri.brighton.ac.uk

Which words are particularly characteristic of a text? A survey of statistical approaches.

Adam Kilgarriff
ITRI
University of Brighton

Abstract

Which words are particularly characteristic of a text, or body of texts? Researchers have wanted to answer this question for a variety of reasons, in a variety of academic disciplines. The paper will survey the answers, comparing the different statistics which have been used and the different circumstances in which they are applicable.

1 Introduction

Which words are particularly characteristic of a text, or body of texts? Researchers have wanted to answer this question for a variety of reasons, in a variety of academic disciplines. The paper surveys answers given in linguistics, social science, information retrieval, natural language processing and speech processing, and considers which are flawed, and under which circumstances in which the remaining ones are applicable.

In Section 2 I set up a simple framework for addressing the question and consider the χ^2-test, the Mann-Whitney test, the t-test, mutual information (MI), the log-likelihood test and TF.IDF, an information retrieval measure, as candidate statistics. $\chi^2$and MI are found to have serious shortcomings. Log-likelihood is suitable where the texts for comparison are viewed as having no internal divisions. Mann-Whitney and TF.IDF are each appropriate under specific circumstances and for specific goals.

In Section 3 I discuss the issue of clumpiness: where a word has occurred in a text, the chances of it occurring again increase. The independence assumption underlying simple stochastic models is unsustainable. More sophisticated, two-parameter models are required, as argued by Church and Gale (1995b).

In Section 4 I consider the route taken in content analysis and the multidimensional approach to language variation. Information on the characteristics of a text is compressed into a small number of summary statistics, which compare the text to independently-generated constructs of the language and are well suited to human interpretation. I finish with some comments on the contrast between high-frequency and low-frequency words.

Measuring the distinctiveness of words in corpora is, for some purposes, homologous to measuring the distinctiveness of combinations of words in a corpus (e.g. bigrams and similar). Daille (1995) presents a review and assessment of measures of strength of co-occurrence, in a paper which can be seen as a complement to this one. She considers a wider range of measures, but her best candidates are considered here.

2 A simple framework

First let us consider the simplest case. We ignore the internal structure of the texts, so the texts could be corpora comprising many documents. For two texts, which words best characterise their differences? For word w in texts X and Y, this might be represented in a contingency table as follows:

	X	Y	
w	a	b	a+b
not w	c	d	c+d
	a+c	b+d	a+b+c+d=N

There are a occurrences of w in text X (which contains $a+c$ words) and b in Y (which has $b+d$ words).[1]

[1] For bigrams, the columns are for w_2 and not-w_2 rather than text X and text Y, and the window of words within which w and w_2 must both occur for it to count as a co-occurrence must also be defined.

2.1 χ^2 test

We now need to relate our question to a hypthesis we can test. The obvious candidate is the null hypothesis that both texts comprise words drawn randomly from some larger population; for a contingency table of dimensions $m \times n$, if the null hypothesis is true, the statistic

$$\Sigma \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$$

(where O is the observed value, E is the expected value calculated on the basis of the joint corpus, and the sum is over the cells of the contingency table) will be χ^2-distributed with $(m-1) \times (n-1)$ degrees of freedom.[2] For our 2×2 contingency table the statistic has one degree of freedom and we apply Yate's correction, subtracting $\frac{1}{2}$ from O-E before squaring. Wherever the statistic is greater than the critical value of 7.88, we conclude with 99.5% confidence that, in terms of the word we are looking at, X and Y are not random samples of the same larger population.

This is the strategy adopted by Hofland and Johansson (1982) to identify where words are more common in British than American English or vice versa. X was the LOB corpus, Y was the Brown, and, in the table where they make the comparison, the χ^2 value for each word is given, with the values starred where they exceeded the critical value so one might infer that the LOB-Brown difference was non-random.

Looking at the LOB-Brown comparison, we find that this is true for very many words, and for almost all very common words. Most of the time, the null hypothesis is defeated. Does this show that all those words have systematically different patterns of usage in British and American English, the two types that the two corpora were designed to represent?

To test this, I took two corpora which were indisputably of the same language type: each was a random subset of the BNC.[3] As in the LOB-Brown comparison, for very many words, including most common words, the null hypothesis was defeated.

This reveals a bald, obvious fact about language. Words are not selected at random. There is no *a priori* reason to expect them to behave as if they had been, and indeed they do not. The LOB-Brown differences cannot in general be interpreted as British-American differences: it is in the nature of language that any two collections of texts, covering a wide range of registers (and comprising, say, less than a thousand samples of over a thousand words each) will show such differences. While it might seem plausible that oddities would in some way balance out to give a population that was indistinguishable from one where the individual words (as opposed to the texts) had been randomly selected, this turns out not to be the case.

Let us look closer at why this occurs. A key word in the last paragraph is 'indistinguishable'. In hypothesis testing, the objective is generally to see if the population can be distinguished from one that has been randomly generated – or, in our case, to see if the two populations are distinguishable from two populations which have been randomly generated on the basis of the frequencies in the joint corpus. Since words in a text are not random, we know that our corpora are not randomly generated. The only question, then, is whether there is enough evidence to say that they are not, with confidence. In general, where a word is more common, there is more evidence. This is why a higher proportion of common words than of rare ones defeat the null hypothesis.

The χ^2-test can be used for all sizes of contingency tables, so can be used to compare two corpora in respect of a set of words, large or small, rather than one-word-at-a-time. In all the experiments in which I have compared corpora in respect of a substantial set of words, the null hypothesis has been defeated (by a huge margin).

The original question was not about which words are random but about which words are most distinctive. It might seem that these are converses, and that the words with the highest values for the χ^2 statistic – those for which the null hypothesis is most soundly defeated – will also be the ones which are most distinctive to one corpus or the other. Where the overall frequency for a word in the joint corpus is held constant, this is valid, but as we have seen, for very common words, high χ^2 values are associ-

[2] Provided all expected values are over a threshold of 5.

[3] The sampling was as follows: all texts shorter than 20,000 words were excluded. Frequency lists were generated for the first 20,000 running words of all others. Half the lists were then randomly assigned to each of two subcorpora. Frequency lists for each corpus were generated. The experiment is fully more described in (Kilgarriff & Salkie, 1996).

ated with the sheer quantity of evidence and are not necessarily associated with a pre-theoretical notion of distinctiveness (and for words with expected frequency less than 5, the test is not usable).

2.2 Mann-Whitney ranks test

An alternative which is suited to the LOB-Brown comparison is the Mann-Whitney (also known as Wilcoxon) ranks test. The LOB and Brown both contain 2,000 word samples, so the numbers of occurrences of a word are directly comparable across all samples in both corpora. The test addresses the null hypothesis – that all samples are from the same population – by seeing whether the counts from the texts in the one corpus are usually bigger than ones from the other, or usually smaller, or similarly spread. The frequencies of word w in each sample are labelled with the corpus they come from, put in rank order, and numbered from 1 to $m+n$ (where there are m samples in the one corpus and n in the other) according to their rank. All the ranks of items coming from the smaller corpus are summed. The sum is then compared with the figure that would be expected on the basis of the null hypothesis, as tabulated in statistics textbooks.

2.3 *t*-test

The (unrelated) t-test, which operates on counts rather than on rank order of counts, could also be used, but the Mann-Whitney test has the advantage of being non-parametric. It makes no assumptions about the data obeying any particular distribution. The t-test is only valid where the data is normally distributed — which is not in general the case for word counts.

2.4 Mutual Information

Another approach uses the Mutual Information (MI) statistic (Church & Hanks, 1989). This simply takes the (log of the) ratio of the word's relative frequency in one corpus to its relative frequency in the joint corpus, so if a word has 50 occurrences in the one corpus, and 7 in another corpus of the same size, the MI for the word and the first corpus is $\log(\frac{50}{1} \times \frac{2}{57})$. In terms of the contingency table above,

$$MI_{w,X} = log_2 \left(\frac{a}{a+c} \times \frac{N}{(a+b)} \right)$$

This is an information theoretic measure as distinct from one based in statistical hypothesis testing, and it makes no reference to hypotheses. Rather, it states how much information word w provides about corpus X, and vice versa. It was introduced into language engineering as a measure for co-occurrence, where it specifies the information one word supplies about another.

Church and Hanks state that MI is invalid for low counts, suggesting a threshold of 5. In contrast to χ^2, there is no notion in MI of evidence accumulating. MI, for our purposes, is a relation between a corpus and a word: if the corpus is held constant, it is usually rare words which give the highest MI. This contrasts with common words tending to have the highest χ^2 scores. Church and Hanks proposed MI as a tool to help lexicographers isolate salient co-occurring terms. Several years on, it is evident that MI overemphasises rare terms, relative to lexicographers' judgements of salience, while χ^2 correspondingly overemphasises common terms.

2.5 Log-likelihood (G^2)

Dunning (1993) is concerned at the invalidity of both χ^2 and MI where counts are low. The word he uses is 'surprise'; he wants to quantify how surprising various events are. He points out that rare events, such as the occurrence of many words and most bigrams in almost any corpus, play a large role in many language engineering tasks yet in these cases both MI and χ^2 statistics are invalid. He then presents the log-likelihood statistic, which gives an accurate measure of how surprising an event is even where it has occurred only once. For our contingency table, it can be calculated as

$$\begin{aligned} G^2 = 2(&a\log(a)+b\log(b)+c\log(c)+d\log(d) \\ &-(a+b)\log(a+b)-(a+c)\log(a+c) \\ &-(b+d)\log(b+d)-(c+d)\log(c+d) \\ &+(a+b+c+d)\log(a+b+c+d)) \end{aligned}$$

Daille (1995) determines empirically that it is an effective measure for finding terms. In relation to our simple case, of finding the most surprisingly frequent words in a corpus without looking at the internal structure of the corpus, G^2 is a mathematically well-grounded and accurate measure of surprisingness, and early indications are that, at least for low and medium frequency words such as those in Daille's study, it corresponds reasonably well to human judgements of distinctiveness.

2.6 Information Retrieval

The question, "Which words are particularly characteristic of a text?" is at the heart of information retrieval (IR). These are the words which will be the most appropriate key words and search terms. The general IR problem is to retrieve just the documents which are relevant to a user's query, from a database of many documents.[4]

A very simple method would be to recall just those documents containing one or more of the search terms. Since the user does not want to be swamped with 'potentially relevant' documents, this method is viable only if none of the search terms occur in many documents. Also, one might want to rank the documents, putting those containing more of the search terms at the top of the list. This suggests two modifications to the very simple method which give us the widely-used TF.IDF statistic (Salton, 1989, p 280, and references therein). Firstly a search term is of more value, the less documents it occurs in: IDF (inverse document frequency) is calculated, for each term in a collection, as the log of the inverse of the proportion of documents it occurs in. Secondly, a term is more likely to be important in a document, the more times it occurs in it: TF for a term and a document is simply the number of times the term occurs in the document.

Now, rather than simply registering a hit if there are any matches between a query term and a term in a document, we accumulate the TF.IDF scores for each match. We can then rank the hits, with the documents with the highest summed TF.IDF coming at the top of the list. This has been found to be a successful approach to retrieval (Robertson & Sparck Jones, 1994).[5]

Two considerations regarding this scheme are:

- As described so far, it does not normalise for document length. In IR applications, TF is usually normalised by the length of the document. The discussion above shows that this is not altogether satisfactory. A single use of a word in a hundred-word document is far less noteworthy than ten uses of the word in a thousand-word document, but, if we normalise TF, they become equivalent.
- Very common words will be present in all documents. In this case, IDF = log1 = 0 and TF.IDF collapses to zero. This point is not of particular interest to IR, as IR generally puts very common words, on a stop list and ignores them, but it is a severe constraint on the generality of TF.IDF.

3 Clumpiness

As Church and Gale (1995b) say, words come in clumps; unlike lightning, they often strike twice. Where a word occurs once in a text, you are substantially more likely to see it again than if it had not occurred once. Once a corpus is seen as having internal structure –that is, comprising distinct texts– the independence assumption is unsustainable. The issue is discussed in detail in Church and Gale (1995b), upon which this section is based (see also Church and Gale (1995a)).

3.1 Probability distributions

The three probability distributions which are most commonly cited in the literature are the poisson, the binomial, and the normal. (Dunning refers to the multinomial, but for current purposes this is equivalent to the binomial.) The normal distribution is most often used as a convenient approximation to the binomial or poisson, where the mean is large, as justified by the Central Limit Theorem. For all three cases (poisson, binomial, or normal approximating to either) the distribution has just one parameter. Mean and variance do not vary independently: for the poisson they are equal, and for the binomial, if the expected value of the mean is p, the expected value of the variance is $p(1-p)$.

To relate this to word-counting, consider the situation in which there are a number of same-length text samples. If words followed a poisson or binomial distribution, then, if a word occurred, on average, c times in a sample, the expected value for the variance of hits-per-sample is also c (or, in the binomial case, slightly less: the difference is negligible for all but the most common words). As various authors have found, this is not the case. Most of the time, the variance is greater than the mean. This was true

[4] We assume full-text searching. Also, issues such as stemming and stop lists are not considered, as they do not directly affect the statistical considerations.

[5] They also provide a 'tuning constant' for adjusting the relative weight given to TF and IDF to optimise performance.

for all but two of the 5,000 most common words in the BNC.[6]

3.2 Poisson mixtures

Following Mosteller and Wallace (1964), Gale and Church identify Poisson models as belonging to the right family of distributions for describing word frequencies, and then generalise so that the single poisson parameter is itself variable and governed by a probability distribution. A 'poisson mixture' distribution can then be designed with parameters set in such a way that, for a word of a given level of clumpiness and overall frequency in the corpus, the theoretical distribution models the number of documents it occurs in and the frequencies it has in those documents.

They list a number of ways in which clumping –or, more technically, 'deviation from poisson'– can be measured. IDF is one, variance another, and they present three more. These empirical measures of clumpiness can then be used to set the second parameter of the poisson-mixture probability model. They show how these improved models can be put to work within a Bayesian approach to author identification.

3.3 Adjusted frequencies

The literature includes some proposals that word counts for a corpus should be adjusted to reflect clumpiness, with a word's frequency being adjusted downwards, the clumpier it is. The issues are described in Francis and Kučera (1982, pp 461–464). Francis and Kučera use a measure they call AF, attributed (indirectly) to J. Lanke of Lund University. It is defined as:

$$\mathrm{AF} = \left(\sum_{i=1}^{n} (d_i x_i)^{\frac{1}{2}} \right)^2$$

where the corpus is divided into n categories (which could be texts but, in Francis and Kučera's analysis, are genres, each of which contain numerous texts); d_i is the proportion of the corpus in that category; and x_i is the count for the word in the category.

Adjusting frequencies is of importance where the rank order is to be used directly for some purpose, for example, for choosing vocabulary for language-teaching, or in other circumstances where a single-parameter account of a word's distribution is wanted. Here, I mention it for purposes of completeness. A two-parameter model as proposed by Church and Gale gives a more accurate picture of a word's behaviour than any one-parameter model.

4 Summary statistics for human interpretation

4.1 Content analysis

Content analysis is the social science tradition of quantitative analysis of texts to determine themes. It was particularly popular in the 1950s and 60s, a landmark being the General Enquirer (Stone, Dunphy, Smith, & Ogilvie, 1966), an early computerised system. Studies using the method have investigated a great range of topics, from analyses of propaganda and of changes in the tone of political communiqués over time, to psychotherapeutic interviews and the social psychology of interactions between management, staff and patients in nursing homes. The approach continues to be taught in some social science 'methods' courses, and to be used in political science (Fan, 1988), psychology (Smith, 1992) and market research (Wilson & Rayson, 1993). The basic method is to:

- identify a set of 'concepts' which words might fall into, on the basis of a theoretical understanding of the situation;
- classify words into these concepts, to give a content analysis dictionary;
- take the texts (these will often be transcribed spoken material);
- for each text, count the number of occurrences of each concept.

One recent scheme, Minnesota Contextual Content Analysis (McTavish & Pirro, 1990, MCCA), uses both a set of 116 concepts and an additional, more general level of 4 'contexts'. Norms for levels of usage of each concept come with the MCCA system, and scores for each concept are defined by taking the difference between the norm and the count for each concept-text pair (and dividing by the standard deviation of the concept across contexts; this normalisation is not discussed in McTavish and Pirro (1990); nor

[6] Figures based on the standard-document-length subset of the BNC described above.

is the source of the norms). The concept scores are then directly comparable, between concepts and between texts. The approach is primarily descriptive: it provides a new way of describing texts, which it is then for the researcher to interpret and explain, so MCCA does nothing more with the concept scores.

It does however also provide the context scores. These serve several purposes, including

> [to] contribute to a kind of "triangulation", which would help to locate any potential text in relation to each of the "marker" contexts. (p 250)

The validity of this kind of analysis is to be found in its predictive power. A content analysis study of open-ended conversations between husbands and wives was able to classify the couples as 'seeking divorce', 'seeking outside help', or 'coping' (McDonald & Weidetke, 1979, quoted in McTavish and Pirro, p 260).

4.2 Multi-dimensional analysis

A major goal of sociolinguistics is to identify the main ways in which language varies, from group to group and context to context. Biber (1988, 1995) identifies the main dimensions of variation for English and three other languages using the following method:

- Gather a set of text samples to cover a wide range of language varieties;
- Enter them ("the corpus") into the computer;
- Identify a set of linguistic features which are likely to serve as discriminators for different varieties;
- Count the number of occurrences of each linguistic feature in each text sample;
- Perform a factor analysis (a statistical procedure) to identify which linguistic features tend to co-occur in texts. The output is a set of "dimensions", each of which carry a weighting for each of the linguistic features.
- Interpret each dimension, to identify what linguistic features, and what corresponding communicative functions, high-positive and high-negative values on the dimension correspond to.

For English, Biber identifies seven dimensions, numbered in decreasing order of significance (so dimension 1 accounts for the largest part of the non-randomness of the data, dimension 2, the next largest, etc.) The first he calls "Involved versus Informational Production". Texts getting high positive scores are typically spoken, typically conversations. Texts getting high negative scores are academic prose and official documents. The linguistic features with the highest positive weightings are "private" verbs (*assume, believe* etc.), *that*-deletion, contractions, present tense verbs, and second person pronouns. The linguistic features with the highest negative weightings are nouns, word length, prepositions, and type-token ratio. The two books cited above present an impressive array of evidence for the explanatory power of the multidimensional approach.

Any text can be given a score for any dimension, by counting the numbers of occurrences of the linguistic features in the text, weighting, and summing. The approach offers the possibility of "triangulation", placing a text within the space of English language variation, in a manner comparable to MCCA's context scores but using linguistic rather than social-science constructs, and using a statistical procedure rather than theory to identify the dimensions.

Approaches discussed in sections 2 and 3 can be viewed as taking each word as defining a distinct dimension of a vector. Biber first reduces the dimensionality of the space to a level where it is manageable by a human, and then offers contrasts between texts, and comments about what is distinctive about a text, in terms of these seven dimensions.[7] He thereby achieves some generalisation: he can describe how classes of features behave (having previously determined that the set of features forms a coherent class), whereas the other methods can only talk about the behaviour of individual features. (Some of Biber's features are words, while others are word-classes and grammatical constructions. Over the coming year I intend to explore the possibility of replicating his approach using only words as features.)

[7] Reducing the dimensionality of the problem has also been explored in IR: see Schütze and Pederson (1995), Dumais, Furnas, Landauer, Deerwester, and Harshman (1988).

5 Discussion

Clearly, people working in the area of measuring what is distinctive about a text have had a variety of goals. Some have been producing figures primarily for further automatic manipulation, others have had human scrutiny in mind. Some have been comparing texts with texts, others, texts or corpora with corpora, and others again have been making comparisons with norms for the language at large. (These issues are considered in more detail in Kilgarriff and Salkie (1996).) Some (Biber, Mosteller and Wallace) have looked more closely at high-frequency, form words; others (McTavish and Pirro, Dunning, Church and Gale) at medium and low frequency words.

The words in a corpus approximate to a Zipfian distribution, in which the product of rank order and frequency is constant. So, to a first approximation, the most common word in a corpus is a hundred times as common as the hundredth most common, a thousand times as common as the thousandth, and a million times as common as the millionth. This is a very skewed distribution. The few very common words have several orders of magnitude more occurrences than most others. The different ends of the range tend to have very different statistical behaviour. Thus, as we have seen, high-frequency words tend to give very high χ_2 scores whereas very high MI scores come from low-frequency words. Variance, as we have seen, is almost always greater than the mean, and the ratio tends to increase with word frequency.

Linguists have long made a distinction approximating to the high/low frequency contrast: form words (or 'grammar words' or 'closed class words') *vs.* content words (or 'lexical words' or 'open class words'). The relation between the distinct linguistic behaviour, and the distinct statistical behaviour of high-frequency words is obvious yet intriguing. It would not be surprising if we cannot find a statistic which works well for both high and medium-to-low frequency words. It is far from clear what a comparison of the distinctiveness of a very common word and a rare word would mean.

Acknowledgments

The work was undertaken under EPSRC grant GR/K/18931. It has benefited from discussions with Ted Dunning and Mark Lauer.

Reference

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions in Register Variation*. Cambridge University Press.

Church, K., & Gale, W. (1995a). Inverse document frequency (IDF): a measure of deviations from Poisson. In Yarowsky, D., & Church, K. (Eds.), *Third Workshop on very large corpora*, pp. 121–130 MIT.

Church, K., & Gale, W. (1995b). Poisson mixtures. *Journal of Natural Language Engineering*, *1*(2), 163–190.

Church, K., & Hanks, P. (1989). Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting*, pp. 76–83 Vancouver.

Daille, B. (1995). Combined approach for terminology extraction: lexical statistics and linguistic filtering. Tech. rep. 5, UCREL, Lancaster University.

Dumais, S., Furnas, G., Landauer, T., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of CHI '88*, pp. 281–285.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Fan, D. P. (1988). *Predictions of public opinion from the mass media : computer content analysis and mathematical modeling.* Greenwood Press, New York.

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage: lexicon and grammar.* Houghton Mifflin.

Hofland, K., & Johansson, S. (Eds.). (1982). *Word Frequencies in British and American English.* The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Kilgarriff, A. (1996). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings, COLING workshop on very large corpora* Copenhagen. submitted.

Kilgarriff, A., & Salkie, R. (1996). Corpus similarity and homogeneity via word frequency. In *EURALEX Proceedings* Göteborg, Sweden.

McDonald, C., & Weidetke, B. (1979). Testing marriage climate. Master's thesis, Iowa State University, Ames, Iowa.

McTavish, D. G., & Pirro, E. B. (1990). Contextual content analysis. *Quality and Quantity*, *24*, 245–265.

Mosteller, F., & Wallace, D. L. (1964). *Applied Bayesian and Classical Inference - The Case of The Federalist Papers.* Springer Series in Satistics, Springer-Verlag.

Robertson, S. E., & Sparck Jones, K. (1994). Simple, proven approaches to text retrieval. Tech. rep. 356, Computer Laboratory, Cambridge University.

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

Schütze, H., & Pederson, J. O. (1995). Information retrieval based on word senses. In *Proceedings, ACM Special Interest Group on Information retrieval.*

Smith, C. P. (1992). *Motivation and personality: handbook of thematic content analysis.*

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Enquirer: A Computer approach to content analysis.* MIT Press, Cambridge, Mass.

Wilson, A., & Rayson, P. (1993). The automatic content analysis of spoken discourse. In *ICAME Proceedings.*