The Statistics of Text: New methods for Content Analysis

Will Lowe
wlowe@latte.harvard.edu
Center for Basic Research in the Social Sciences
Harvard University

Computer content analysis (CCA) is used across the social sciences, and is beginning to find a range of applications in political science. These have traditionally been concentrated on political communication and policy analysis in America and Western Europe (Laver and Garry, 2000; Pennings and Keman, 2002), although CCA is potentially appropriate anywhere traditional discourse analysis might normally be considered (Neuendorf, 2002; Abdelal et al., 2003). In the dominant approach to CCA, the researcher constructs a category system or 'dictionary' that associates a set of words with each theoretically relevant concept, and summarizes a document's content in a vector of category occurrence frequencies. More linguistically sophisticated methods have been used for particular research problems; two important examples are the use of partial parsing and information extraction technology for events data in the KEDS project, (Schrodt, 1994; Gerner et al., 1994) and Young's software for inferring decision maker's cognitive maps in political psychology (Young, 1996).

Methods that look for keywords, and methods that construct detailed graph structures and relational maps are two extremes of CCA. The first makes almost no assumptions about how text is actually generated, since it is based on a theory of keyword content discussed below. The latter is based on a highly developed theory of how e.g. causal relationships are expressed in text. Since there is no free lunch in data analysis, the latter methods are highly specific to particular genres of text in particular languages. KEDS is optimized for Reuters news feed, which has a fairly stereotypical structure, and Young's software is designed for political speeches. Both require data in, or translated into English. Dictionary methods in contrast, require only a time series with a finite number of possible discrete events, an assumption that not even specific to linguistic data. There is in principle a continuum of methods between these levels of complexity, though in practice these have been the dominant two options.

This paper has two parts. The first part describes some new complementary exploratory methods that are not keyword based, but also do not require strong assumptions about how concepts are expressed in text. They are a first attempt to move some of the ideas behind discourse analysis and cognitive map construction closer to textual data. The methods are not language or content specific. Consequently they cannot be as targeted as a customized application such as KEDS, but hopefully can provide useful exploratory analyses when a more complex method is not available,

or not yet developed.

Like all analytical tools, CCA methods embody determinate, though often unarticulated assumptions about the structure of text. The second part of the paper attempts to situate the new methods among existing approaches to CCA. This section considers each CCA method's implicit assumptions about the data generating mechanism for text, and asks under what circumstances each might be expected to be appropriate. Presently content analysis stands somewhat apart from the body of political methodology, presumably because of its roots in interpretative methods, but it is as essential to ask the same questions of a method for inferring content as any other kind of estimator. This section is therefore intended as a first small step towards a general treatment.

Development of these methods is part of ongoing work on the Identity Project, a Weatherhead Center for International Affairs project to quantify politically relevant senses of the concept of identity through various forms of text analysis[1]. The research interests of project members (Russia, post-Soviet States, and China) require the ability to deal with other languages. Readers may be interested in one of the more practical results of the project, a free open-source multilingual content analysis program that runs on all operating systems and provides most of the functions of commercial offerings[2].

1 Measures of Contextual Similarity

Standard approaches to CCA require a dictionary and are appropriate to almost any type of text, provided that the purpose of analysis is to confirm a hypothesis. Dictionary-based methods are essentially confirmatory because the dictionary contains a set of categories that are theoretically important[3]. If what is theoretically important is not yet known, the only other option is to use somebody else's category set (e.g. one of those described in Pennebacker and King, 1999; Stone, 1997; Hart, 1997). This section is an attempt to describe measures of content that do not assume a developed theory, and concentrate on *word usage* as a guide to content.

The importance of word use for understanding meaning was first pursued in the philosophy of language (Wittgenstein, 1958; Quine, 1960, 1961) and linguistics (Harris, 1954, 1963b; Cruse, 1986). More recently, cognitive scientists have provided computational treatments Redington and Chater (1997); Lowe (2001); Landauer and Dumais (1997). These 'contextual' approaches to meaning assume that word's meaning is constituted primarily by its use rather than, e.g. its reference. Specifically, it is constituted by constraints on the linguistic contexts a word can appear in. In particular, a contextual theory of meaning states that two words are similar in meaning to the

[1]Originally this paper was to demonstrate the new methods on a sequence of articles in publications from Goskomstat, the Russian government's economic body, as the country transitioned to a market economy. Regrettably, data problems made this impossible to achieve in time for the conference.

[2]The current version runs in Java, and is available at http://www.people.fas.harvard.edu/∼wlowe/CCA.html The next version release (late April) will contain contextual similarity functions. Before then, these functions are available from the author on request.

[3]This essentially holds for automated dictionary construction methods that use previously content-analysed documents - the categories or dimensions of the previous analysis are presupposed.

side with israel in peace and	security	like all other people palestinians deserve
conflict above all our principles and our	security	are challenged today by outlaw groups
people of iraq deserve it the	security	of all nations requires it. Free
the just demands of peace and	security	will be met or action will

Table 1: Concordance with a 6 word window, taken from G. W. Bush's address to the United Nations, September 12th, 2002

extent that they can be substituted for each other in the same context. Equivalently: two words are similar in meaning to the extent that are talked about in the same way, or share similar linguistic contexts[4]. Statistical models of this theory concentrate on quantifying 'spoken about in the same way', and 'occurs in similar contexts'.

In political science, contextual approaches are an attempt to describe the meaning of a word *to* some author or audience, in a particular discourse, in a quantitative and replicable way. Although this view of meaning may appear to be hopelessly tied to interpretation, computational approaches to contexts can also be seen as the first steps toward the automated construction of the cognitive maps (Axelrod, 1976) that are used in political psychology and international relations (Johnston, 1996).

Contextual methods can also be usefully seen as a way of quantifying a standard concordance analysis. Concordances, or keywords-in-context, provide a direct way to examine word usages in a document by extracting a word all the instances of a term. For example, Table 1 is the following is a subset of lines from a concordance for the word 'security' in a speech by G. W. Bush.

A concordance analysis might manually compare these entries to the same word in other speeches and to other audiences, or to the concordances of terms denoting the use of force, to see how similar these are. Statistical measures of contextual similarity can be seen as a way to quantify how similar these contexts are to those of other words, without explicitly generating concordance and examining them, and with a quantitative measure of similarity.

With the concordance in mind, it might seem that an obvious direct way to encode the contexts enumerated in Table 1 would be to construct a vector of word counts, one for each 'context' word that appears around the 'target' word 'security'. We might then examine distances between vectors, or cluster analysis as a measure of contextual similarity. Unfortunately, although the idea is sound, the statistics of text make using counts a bad idea.

[4]This substitution model of meaning seems to be presupposed by the GRE, SAT and TOEFL's sentence completion tests, in which candidates' linguistic competence is judged by their ability to pick the most appropriate of several possible completions.

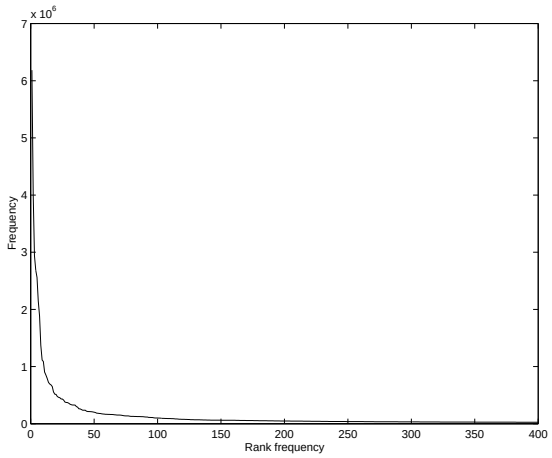

Figure 1: Occurrence frequency plotted against rank in a frequency list for the 400 most frequent word stems in the English.

1.1 The Statistics of Text

Text is the basic and most important source of material in political science. But although this is extremely obvious, it is striking that a vast amount of statistical work proceeds without using it. This may be due to the absence of adequate understanding of the psychology of interpretation. But there is a more mundane reason statistical text analysis is hard: As objects of statistical inquiry, words are just *awkward*.

Word frequencies, the basis of all content analysis procedures, are distributed according to Zipf's Law (Zipf, 1949; Mandelbrot, 1954; Li, 1992) which relates the probability π of a word to its rank r in a ranked frequency list as $\pi \sim 1/r^{\alpha}$ with α close to unity. Similar distributions appear in biology, econometrics and computer science (see Mitzenmacher, 2001, for a review). Figure 1 shows the empirical frequency distribution for the 400 most frequent word stems in the British National Corpus, a 100 million word corpus of British English. Essentially identical plots can be constructed for American English, and indeed all known languages.

A handful of word types dominate this list: the 10 most frequent word stems in the BNC are 'the', 'be', 'of', 'and', 'to', 'a', 'in', 'have', 'that' and 'it', constituting slightly over one quarter of all tokens in the corpus[5]. These words are very frequent but relatively contentless. As is often the case in the social sciences, the most interesting events occur relatively rarely.

From a statistical perspective the power scaling of Zipf's law ensures that the majority of words occur very infrequently and generate a severe sparse data problem. Another problem is that techniques that assume Normal or near Normal distributions are seldom directly applicable. Word frequencies are count data with highly skewed marginals.

The statistics of text mean that the obvious, direct approach to representing context is highly

[5]In this corpus, 25974687 / 99985962 $\approx$ 0.26

	t	$\neg t$	
c	$f_{\langle c\ t\rangle}$	$f_{\langle c\ \neg t\rangle}$	$f_{\langle c\rangle}$
$\neg c$	$f_{\langle \neg c\ t\rangle}$	$f_{\langle \neg c\ \neg t\rangle}$	$f_{\langle \neg c\rangle}$
	$f_{\langle t\rangle}$	$f_{\langle \neg t\rangle}$	N

Table 2: Contingency table for words c and t. $f_{\langle c\ t\rangle}$ denotes the number of times c occurs one word to the left of t.

biased. This is because word choices are not purely driven by content; many word appearances in the window occur due to the syntactic constraints of making a grammatical sentence, or simply to chance.

As an example of the problem, consider two words t and c with occurrence probabilities π_t and π_c in a text containing N words. If t and c have *no* relation to each other then we can model their empirical frequencies as $f_{\langle t\rangle} \sim \text{Binomial}(\pi_t, N)$ and $f_{\langle c\rangle} \sim \text{Binomial}(\pi_c, N)$[6]. In this idealisation where t and c are perfectly independent, the expected frequency of the word pair $\langle c\ t\rangle$ (that is, the event where c occurs before t in text) is $N\pi_c\pi_t$, which is linear in the probability of t. Thus when there is no contentful connection between t and c, chance will ensure that they co-occur at a variable rate that depends on their marginal probabilities.

Even when they are related, the variation in the cooccurrence count $f_{\langle c\ t\rangle}$ will be driven by the marginal frequency of the target word as well as its true level of association with c, and the two are not distinguishable. In particular, targets with the same marginal probability will get more similar counts, irrespective of their content relationship to c. This might not be a major problem in domains where marginal frequencies do not vary much, but Zipf's law suggests that language is not one.

These observations suggests that it is the *association* between c and t that is important to quantifying a context, not just the number of times they share one. If vectors are to be good estimators of context, they require a better measure of association.

1.2 A better measure of association

Any measure of association between a target and context word ought to correct for marginal probability differences. One such measure is the odds ratio, as used in contingency tables. To motivate the odds ratio as a measure of association consider the case where the window over which context words are to be counted extends only one word to the left of the target (a degenerate concordance that extends only one word in one direction). We can measure the level of association between t and c, taking into account marginal probabilities by constructing the contingency Table 2. In this table "$\neg t$" denotes any word that is not the t, and N is the number of words in the text. The maximum likelihood estimate of the odds ratio (Bishop et al., 1975) between t and c provides a measure

[6]In fact they will always be slightly distributionally related through their syntactic properties e.g. by the fact that they are both nouns, but we ignore this.

of association corrected for chance:

$$OR(c\,t) = \frac{\widehat{\pi_{c\,t}}/\widehat{\pi_{c\,\neg t}}}{\widehat{\pi_{\neg c\,t}}/\widehat{\pi_{\neg c\neg t}}} = \frac{f_{\langle ct\rangle}/f_{\langle c\neg t\rangle}}{f_{\langle \neg ct\rangle}/f_{\langle \neg c\neg t\rangle}} = \frac{f_{\langle ct\rangle}\ f_{\langle \neg c\neg t\rangle}}{f_{\langle c\neg t\rangle}\ f_{\langle \neg ct\rangle}} \tag{1}$$

To extend this measure to a window of surrounding context words we can imagine constructing a three way table with position relative to the target word as the third dimension. For example, when the window extends two words either side, we would construct four tables with top left entries that are the frequencies of: $\langle c\ *\ t\rangle$, $\langle c\,t\rangle$, $\langle t\,c\rangle$, and $\langle t\ *\ c\rangle$, where the asterisk indicates the occurrence of any word. Assuming that there is no important content information provided by the *exact* position of c's occurrence in the window, we could then collapse the table across position to recover a two by two table. We then compute the odds ratio as before.

We only need imagine this process because it justifies the much more computationally straightforward process of constructing the final table. The only statistics necessary to construct the final table are

- $f_{\langle c\ \ t\rangle}$, to construct the top left hand entry of the collapsed table
- $f_{\langle c\rangle}$ and $f_{\langle t\rangle}$, to construct the top right and bottom left entries in the sub-tables (e.g. $f_{\langle c\ \ \neg t\rangle} = f_{\langle c\rangle} - f_{\langle c\ \ t\rangle}$)
- N, to avoid computing any marginals for quantities involving '$\neg$' (e.g. $f_{\langle \neg c\rangle} = N - f_{\langle c\rangle}$).

Finally, to obtain a symmetrical Normally distributed measure of association we compute $\log\ OR(c\,t)$. Context vectors are thus populated with log odds-ratios rather than counts.

As an example of the difference the different association measure makes, logged odds ratios for the vector corresponding to the word 'security' in a collection of Bush's public addresses[7] since February 2000 correlate with the corresponding cooccurrences only to degree 0.52.

2 Choosing context words

Which words should be used to make up the vectors that describe a word's context? It is tempting to include all the words in the document, so as to be sure not to miss any more contextual detail than necessary. However, the choice of context words is an example of the tradeoff between bias and variance. If highly frequent words are chosen, e.g. the top 10 most frequent, then when their odds ratios (or counts) are used to construct a vector, that vector will probably not describe the target word's context well. Words like 'the', 'be' and 'of' occur about equally around *any* word in English, and although estimates of their degree of association are likely to be very accurate, their ability to represent any context that *discriminates* target words will be small. That is: high

[7]Not including press conferences, there are six speeches in this data set, for the time period between 09/12/2002 (speech to the U.N.) and 02/26/2003 (speech to the American Enterprise Institute.) Available formatted from the author, or from http://www.whitehouse.gov

frequency context words provide a high bias, low variance estimator of context. If more clearly contentful words (that are typically rather low frequency) are chosen to represent context, they are likely to be very informative, but estimates of their levels of association will be unreliable due to sparse data. Low frequency context words make for a low bias but high variance estimator of context.

Unfortunately the optimal balance between bias and variance will be unknown and choice of context words proceeds empirically. In practice however, including high frequency words is less problematic because the odds measure will provide very accurate estimates of very small amounts of association which will be mostly shared by all target words. Consequently, distance measures between target vectors will not be much affected by these multiple small elements.

2.1 Contextual Similarity

When context vectors have been constructed they can be visualized using multidimensional scaling or cluster analysis. If a distance measure between vectors is chosen (euclidean distance is a straightforward choice, though there are reasons to consider the correlation cooefficient) then the distances themselves can be worked with directly. These are then measures of contextual similarity between any two terms in a document[8].

2.2 Examples

This section considers a simple application of contextual similarity measures to Reuters reports on the Bosnian conflict, and to Tony Blair's speeches on Iraq. The examples are intended primarily to demonstrate the face validity of structures generated by contextual similarity measures. More detailed work related to the Identity Project is in progress[9].

Figure 2 is a dendrogram representing the contextual similarities of the 300 most frequent nouns in Reuters coverage of the Bosnian conflict 1993-94. Context words were chosen to be all words that occurred more than 30 times. This structure captures a number of useful thematic connections in this conflict, some of which are described below.

The figure plots the full dendoragram for a year of leads in the background and boxes selected subtrees in the foreground. The first point about contextual similarity structure is that it reliably recovers proper names, e.g. *Madeleine Albright* is identified as an *ambassador*. Next Islamic countries diplomatic and military roles in Bosnia are represented by a single tree containing one subbranch for Bosnia-*Herzegovina*, *Malaysia*, *diplomat*, *Iran*, *Indonesia* and *Morocco*, and another

[8]Currently it is not clear how best to think about the sampling properties of contextual measures. One possible approach to standard errors for similarity measures is to make use of the easily computed asymptotic s.e. for log odds ratios in conjunction with an assumption that context word occurrences are conditionally independent given the target. This is future research.

[9]This section was initially intended to show the use of contextual similarity measures on a corpus of Russian text, specifically editorials and invited pieces from the journal of Goskomstat, the State Committee of the Russian Federation on Statistics (http://www.gks.ru), as the organization transitioned to a market economy. Unfortunately due to data problems, this work was not ready for the Midwest meeting.

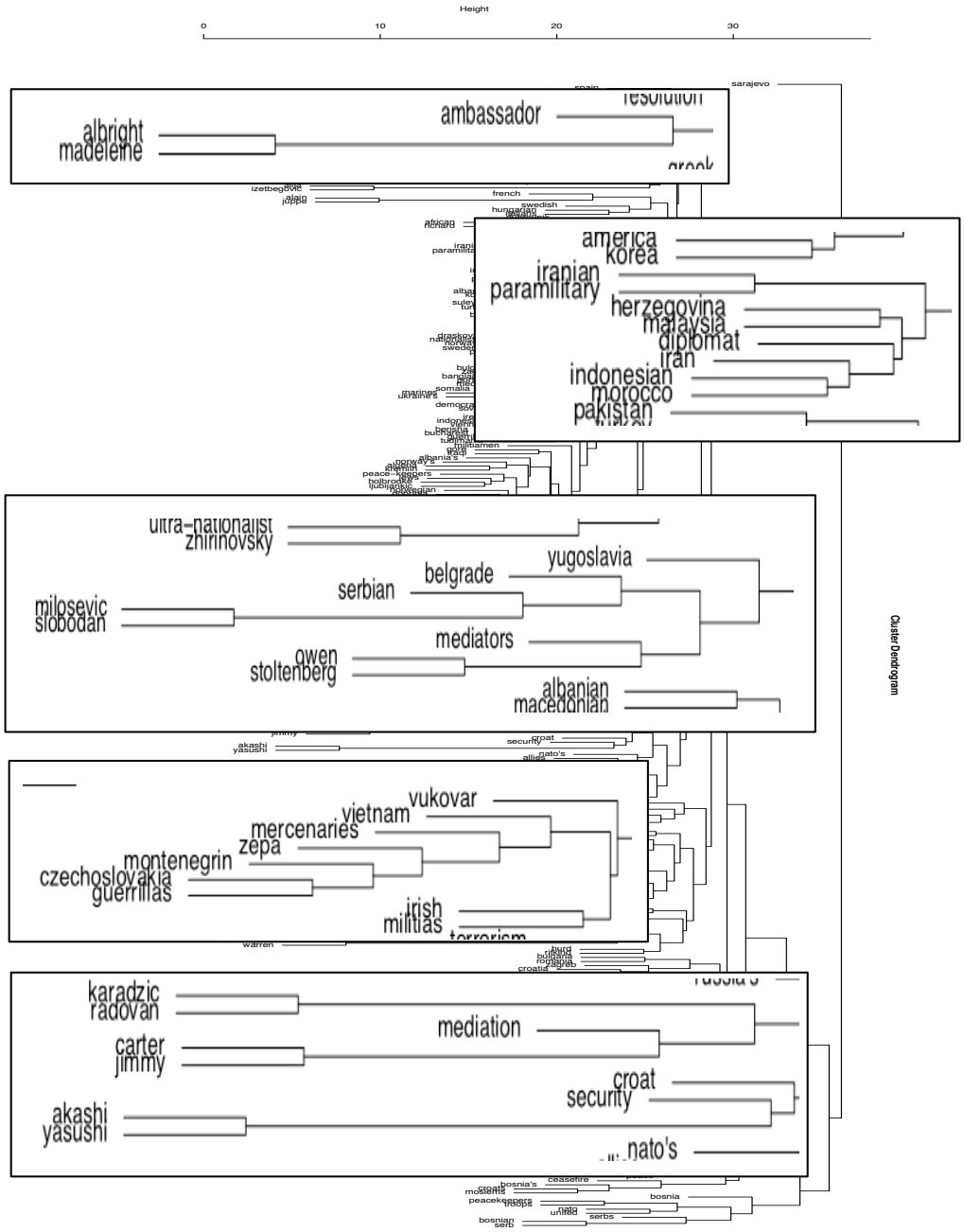

Figure 2: Contextual similarity dendrogram highlights for one year of Reuters newsleads on the Bosnia conflict.

for *Iranian paramilitary* forces. The close grouping of Islamic countries suggests that Reuters reporters are treating them as nearly equivalent. A full subtree connects *Yugoslavia, Belgrade*, and *Serbian*s with *Slobodan Milosevic*, and mediation attempts by *Owen Stoltenberg*. Another subtree is devoted to guerrilla warfare around *Vukovar* and *Zepa*; one branch covers similar themes of *mercenaries* and *guerrillas* in *Vietnam* whilst the other has *Irish* and *militia*. These distances reflect the conceptual similarities invoked by the news coverage. Finally the activities of the Bosnian Serb leader *Radovan Karadic* are grouped with *mediation* attempts by *Jimmy Carter* and *Yasushi Akashi*'s related efforts in Croatia.

The same data using cooccurrence counts rather than log odds ratios is largely uninterpretable, save for proper names which are mostly recovered. The aim of this section is to demonstrate that the new methods have reasonable face validity, and can capture some essential thematic relationships. The results appear to be robust to parameter changes (e.g. in window sizes other than 10, context word frequency cutoffs other than 30 etc.) although further empirical investigation is clearly necessary.

Figure 3 shows highlights from a contextual similarity dendrogram for a year of Tony Blair's speeches about Iraq prior to the beginning of war[10]. The first box shows the familiar thematic map of Blair's stated concerns about Iraq's relation to the United Nations, and the development of weapons of mass destruction. The second box describes Britain's relationship to the rest of the world in liberal terms - *stability, respect, trade* and other *human values*. The next subtree shows *America* and *Europe* as nearly equivalent. From looking at the text corpus this appears to reflect Blair's emphasis on shared values that unite both regions rather than this disagreements. The next two boxes concern Middle Eastern terrorism and the Arab Israeli conflict. The first emphasizes the role of the *security council* in mediating between the *Arab world* and *Israel*, and discusses the idea of a *viable Palestinian state*. In contrast, the box below is the counterpart of this subtree in a dendrogram of Bush's speeches of the same period. Here the discussion is a lot sparser and notes only that *Palestine* be *peaceful* and *democratic*. This contrast appears to be consistent with popular understanding of the difference between the two leader's views on the likely resolution of Arab Israeli conflict.

2.3 Extensions

Although these methods have been derived for words and their contexts, nothing depends on using words as the unit of analysis. For example, if a dictionary is provided then all the words in a particular category can be identified with the category label, and category's contextual similarity can be investigated, either in terms of other words, or entirely in terms of other categories. Indeed, this may alleviate the sparse data problem considerably, though at the cost of requiring a dictionary.

In the second part of this paper we consider how contextual similarity measures fit into the range of existing CCA models.

[10] Available formatted from the author or from http://www.number-10.gov.uk/output/page5.asp.

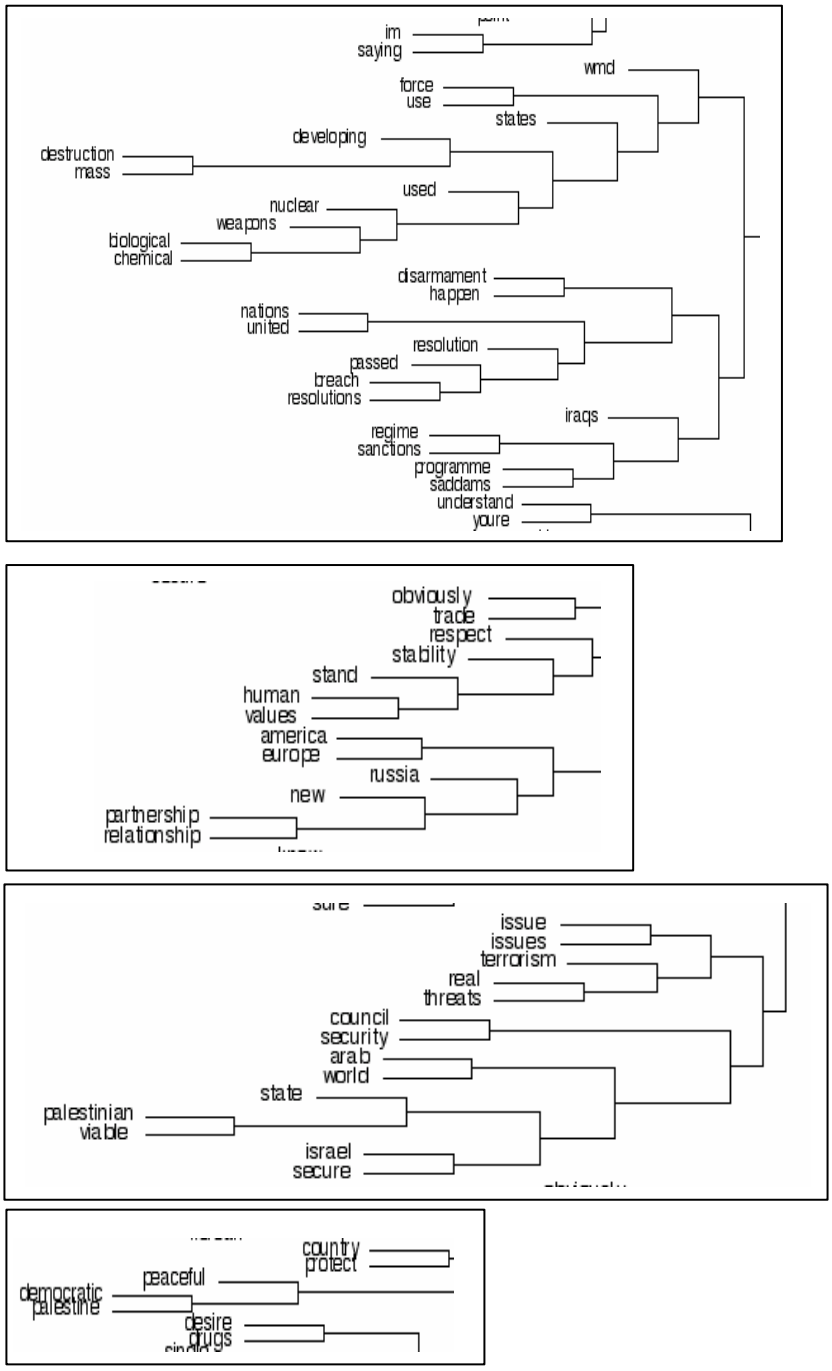

Figure 3: Contextual similarity dendrogram highlights for one year of Blair speeches on Iraq. The bottom box is from Bush speeches in the same period.

3 Data generating mechanisms: A Typology of CCA methods

All content analysis assumes a latent variable model (Everitt, 1984) in which the theoretically relevant content, e.g. the policy position expressed in a manifesto, or the degree of anti-semitism in a speech, is unobserved. Words, phrases or other linguistic structure are then noisy guides to the nature of the content.

This section investigates the circumstances under which different CCA methods are appropriate by taking the statistical approach of considering what sort of data generating mechanisms they presume, and concludes by discussing why contextual similarity methods might apply under several different generation assumptions.

3.1 Dictionary-based CCA

How would text, or speech have to be generated such that a dictionary-based content analysis would be an appropriate way to infer its unobserved content?

Dictionary-based CCA treats words as the relevant unit of observation, and vocabulary choices as the primary indicator of latent content, disregarding any higher level linguistic structure. For dictionary based approaches the data generating model is one where the author has a set of categories to express in a text, an idea of how often she would like to express them, and a known distribution over vocabulary words given category. Dictionary-based content analysis is appropriate to this situation, assuming that the categories a researcher constructs coincide well with those of the author.

Although certainly counterintuitive, this picture of language production is not unreasonable for public discourse. In particular for much political text, particularly manifestos and speeches, there is often a conscious effort on the part of the speaker or her speechwriters to construct a speech that fulfills the assumptions of dictionary-based content analysis rather better than normal text. One reason is that the speaker cannot make strong assumptions about the extent to which the audience shares her mapping from concepts onto particular productions, so it is prudent to ensure to rely most heavily on the lowest level of content that is most likely to be shared. That level typically consists in keywords and simple and iconic imagery. Consistent with this theory, terms may be used in ways more stereotypical than their regular applications in an attempt to ease an unfamiliar listeners inference task. Laver and Garry (2000) noted that when the word 'tax' appeared in British party political manifestos, it was reliably embedded in a sentence that discussed *lowering* taxes. The word tax is then a better indicator of policy position than its literal meaning would suggest.

Constructing a dictionary is a time-consuming and difficult process; the researcher must be reasonably sure that the categories she extracts coincide with those intended by the author, or more importantly, the ones that are prevalent in typical listeners. In part because of the generic difficulty of this process, content analysis dictionaries do not explicitly model p(word — content category), but rather provide for each category a set of words, all of which are presumed to be equally diagnostic.

3.2 Parsing approaches to CCA

The more complex content analysis methods are also relational rather than thematic, but the relations are no longer word associations or cooccurences, but substantively interpretable structures. For example, the KEDS verb template "`Accus + of {plotting | concealing}`" matches sentences of the form: "... accused ... ... of plotting ..."[11]. It is the match to a multi-word template, that is: the exact relation of these words together, that indicates the presence of the content WEIS code 120 (McClelland, 1978). This is a relational rather than thematic approach because the main verb 'conceal' can according to this dictionary equally indicate the content WEIS code 141 in other contexts. In Young's program, seemingly contentless words such as 'since', 'because' and 'if... then' pairings can also be used, for example to extract logical or causal structures.

The data generating process for Young's software is difficult to discover, partly because the software is proprietary, so this discussion concentrates on KEDS. KEDS implements a partial parser (Manning and Schütze, 2000); templates are matched against input text until some combination accounts for enough of the words to assign content. The newswire reporter's mechanism for generating a story is then, according to KEDS, to pick a content category - say the event denoted by WEIS code 120, search through a list of possible stereotypical linguistic constructions, find the one above, and proceed to fill in the gaps with relevant actors. This is a rather plausible data generating mechanism for a reporter with a tight deadline and Reuters' strong constraints over report structure[12].

KEDS is a deterministic parser and does not, in company with most information extraction tools, have an explicit representation of p(sentence — content category)[13]. It is, however, possible to characterize the bias and variance properties of a deterministic parser experimentally (see King and Lowe, 2002, for an example).

3.3 Contextual Similarity Measures

How would text have to be generated for contextual similarity measures to be appropriate to it? The constraints on generating mechanism are clearly weaker for these methods since *any* words could be used to indicate content, and it is only the statistical relationships between their occurrences that are important to the method. An author would have an representation of the relationship between various concepts (an internal cognitive map), and proceed to generate words with the constraint that the elements of the map - countries, concepts, or people - appear surrounded by similar sets of words. This is again an unintuitive picture of text generation, but in important ways a much less limiting one than those described above.

[11] Line 37 of Balkans2001.VERBS, a coding dictionary for the Balkans conflict, available at `http://www.ukans.edu/~keds/data.html`

[12] Reuters reporters are required to construct reports where the first sentence provides all the information in the subsequent article, the second and subsequent sentences elaborate on some clause of the first, and so on in a hierarchical manner (P. Schrodt pers. comm.) The necessary compression of the first sentence makes text more than usually amenable to a template style analysis. It almost certainly would't work with a novel.

[13] The nearest stochastic parsing analogue would be a branching process (Harris, 1963a).

A useful way to understand these surrounding word constraints is to consider what would have to be the case for the method to completely *fail*. One way to complete failure would be to speak so that substantially relevant terms were surrounded by other words at random rates. But this would be word salad, since it would not even be grammatically correct. A less drastic failure would be to have syntactically sound but 'content-empty' text. However this would necessarily fail to express anything about the author's underlying representations, since she would not be able to talk about two similar topics 'in the same way' e.g. apply the same adjectives or use the same constructions on similar terms. It is this notion of 'talking about in the same way' that moves contextual similarity measures towards some of the functions of conventional discourse analysis.

The constraints on data generating process for contextual similarity are relational rather than thematic (Pennings and Keman, 2002). Dictionary-methods have the advantage that they are purely thematic so their data generating mechanisms are fairly easy to express (e.g. generate a word from category A 12 times, category B once etc.) Pure relational constraints are not straightforward to express in an explicit model because they involve a large number of interlocking constraints. But although dictionary-based methods and more complex parsing approaches appear to be assuming rather different data generating mechanisms, contextual similarity measures may in fact be appropriate to both situations.

Similarity and dictionaries

Contextual similarity methods explicitly quantify similarity, but dictionaries also *implicitly* provide a similarity space for words, by grouping them into semantic categories (all members of a category are equally similar). The analogy to contextual methods is better in recent work where words are assigned a degree or probability of category membership, so each word is given a vector of real numbers. As an example of the latter, Pennings and Keman (2002), and Benoit and Laver (2003) have independently shown how to use a previously content analyzed collection of reference documents to induce the probability distribution of p(content category — word). The essence of their approach is to compute, for each word in the reference documents and each category, the probability that that word belongs to a document of each category [14]. This provides a distribution of unobserved categories or positions conditional on words. Probabilities are assigned by computing the proportion of documents of each category or position that contain that word.

Note that although both authors emphasize the probability of content given word, by Bayes Theorem $p(\text{content} \mid \text{word}) \propto p(\text{word} \mid \text{content})$ when different content categories are equiprobable, so this method is simultaneously a way to implicitly provide probability distributions over words given content, a process that is normally too time consuming to do manually, and a way to estimate these probabilities efficiently using a training set rather than the researcher's intuition.

Document level policy positions can be computed by averaging over the position probabilities of each word in the document[15]. But if document level policy positions are *not* computed,

[14] This is a loose summary intended only to emphasize the basic features and similarity between the two approaches. For more details, see Pennings and Keman (2002), and Benoit and Laver (2003).

[15] This procedure is widely used in 'Bayesian' email spam filtering and document classification (Sahami et al., 1998).

- * UP_SECURITY	[170]
- SAID * SECURITY	[051]
- MOUNTED SECURITY *	[182]
- * BOOST + SECURITY	[051]
- * SECURITY_MEASURES	[170]
- * SECURITY_AT	[072]
- SAID WOULD * NECESSARY STEPS TO ENSURE SECURITY	[171]

Table 3: Selected dictionary entries for the word 'security' and their numerical content codes. From KEDS 2001 dictionary on the Balkans c.f. Table 1

the probability distribution over categories for each word itself generates a similarity space; the words themselves can be plotted, scaled or otherwise used in the same way contextual similarity measurements would be.

The difference between these measures and a contextual similarity measure over the same words, is that Laver and Penning's numbers reflect similarity *modulo* the policy scheme or category system; contextual similarity measures remain agnostic about any underlying category scheme, and are consequently more flexible but less targeted.

Similarity and Parsing Methods

Contextual measures may also be appropriate if data is generated consistent with KEDS's assumptions about language because KEDS's dictionary entries are easily interpretable as constraints on context. For example the entries in Table 3 are a randomly chosen selection of KEDS dictionary entries searched for by the keyword 'SECURITY': Jointly, the entries enumerate all possible (or relevant) contexts for the word 'security'. Instances of these templates in text are exactly what contextual similarity measures are extracting and scaling.

Contextual measures are necessarily less informative than the set of dictionary entries in that that may not distinguish between international security and social security because all context words are treated equally, or between military intelligence and the intelligence of the military because they ignore word order. However they can pick up the large scale outlines of word use, without the painstaking construction of this kind of dictionary first.

These observations suggest that, if KEDS were to be used to generate, rather than recognize content, then the sentences it would produce would be good candidates for a contextual similarity analysis. Consequently, even when the data generating mechanism underlying a text is similarly complex, we might expect contextual similarity measures to be useful.

Another way to interpret the same point is to note that if the data generating mechanism that underlies a text is structurally similar to a reasonably complex model such as KEDS, then it is the structure of the model that explains why the contextual similarity model works. On this in-

terpretation the appropriateness of computing a contextual similarity measure doesn't depend on any particular generating model, since many kinds will generate data it can work with. Similarity structures, cognitive maps etc. can then be thought of byproducts of more generic generation mechanisms.

4 Conclusion

This paper has presented some new statistical methods for content analysis that attempt to quantify contextual similarity. It has also shown how they relate to existing methods by asking what assumptions different content analysis procedures make about how text is generated.

Software is available for the new methods from the author, and will be incorporated into the Identity Project's general purpose multilingual content analysis software in its next release in May 2003. Hopefully, some new methods and a free tool will stimulate more statistically grounded content analysis, and encourage methodologists to turn their attention to the important and understudied problems of analyzing text.

References

Abdelal, R., Herrera, Y. M., Johnston, A. I., and McDermott, R. (2003). Identity as a variable (manuscript).

Axelrod, R. (1976). *The Structure of Decision*. Princeton University Press, Princeton NJ.

Benoit, K. and Laver, M. (2003). Estimating Irish party positions using computer wordscoring: The 2002 elections. *Irish Political Studies*, 17(2).

Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge MA.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall, London.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., and Weddle, J. L. (1994). The analysis of political events using machine coded data. *International Studies Quarterly*, 38(1):91–119.

Harris, T. E. (1963a). *The Theory of Branching Processes*. Springer, Berlin, Germany.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2):146–62. Reprinted in J. A. Fodor and J. J. Katz (eds.) The Structure of Language: Readings in the Philosophy of Language, Prentice-Hall NJ.

Harris, Z. S. (1963b). *Structural Linguistics*. University of Chicago Press, Chicago IL.

Hart, R. P. (1997). *DICTION 4.0: The Text Analysis Program*. Sage, Thousand Oaks CA.

Johnston, A. I. (1996). Cultural realism and strategy in Maoist China. In Katzenstein, P. J., editor, *The Culture of National Security*, chapter 7, pages 216–268. Columbia University Press, New York.

King, G. and Lowe, W. (2002). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. Submitted.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, (104):211–240.

Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 44(3):619–634.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Lowe, W. (2001). Towards a theory of semantic space. In Moore, J. D. and Stenning, K., editors, *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 576–581, Mahwah NJ. Lawrence Erlbaum Associates.

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word*, (10):1–27.

Manning, C. D. and Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA.

McClelland, C. (1978). World event/interaction survey, 1966-1978. WEIS Codebook ICPSR 5211, Inter-university consortium for political and social research, University of Southern California.

Mitzenmacher, M. (2001). A brief history of generative models for power law and lognormal distributions. In *Proceedings of the 39th Annual Allerton Conference on Communication, Control and Computing*, pages 182–191. Available at http://www.eecs.harvard.edu/∼michaelm/NEWWORK/papers.html.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Sage, Thousand Oaks CA.

Pennebacker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, (77):1296–1312.

Pennings, P. and Keman, H. (2002). Towards a new methodology of estimating party policy positions. *Quantity and Quality*, 36:55–79.

Quine, W. V. O. (1960). *Word and Object*. MIT Press, Cambridge MA.

Quine, W. V. O. (1961). *From a Logical Point of View*. Harvard University Press, Cambridge MA, second edition.

Redington, M. and Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in the Cognitive Sciences*, 1(7).

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). Bayesian approach to filtering junk email. In *AAAI Workshop on Learning for Text Categorization*, Madison WI.

Schrodt, P. A. (1994). Event data in foreign policy analysis. In Neack, L., Haney, P. J., and Hay, J. A. K., editors, *Foreign Policy Analysis: Continuity and Change in its Second Generation*, pages 145–166. Prentice-Hall, New York.

Stone, P. J. (1997). Thematic text analysis: New agendas for analyzing text content. In Roberts, C., editor, *Text Analysis for the Social Sciences*. Lawrence Erlbaum Associates.

Wittgenstein, L. (1958). *Philosophical Investigations*. Blackwell, Oxford. (trans. G. E. M. Anscombe).

Young, M. (1996). Cognitive mapping meets semantic networks. *Journal of Conflict Resolution*, 40(3):395–414.

Zipf, G. K. (1949). *Human Behavior and the Principal of Least Effort*. Addison Wesley, Reading MA.