April 1993

Towards an Integration of Content Analysis and Discourse Analysis: The Automatic Linkage of Key Relations in Text*

Andrew Wilson

I. Background

Content analysis is a firmly established technique for textual data analysis. In particular, the notion of fully automatic, or at least computer assisted, content analysis has remained a desideratum among researchers working with large bodies of text, with the development of systems such as General Inquirer (Stone et al., 1966), Words (Iker and Harway, 1969), Quester (Cleveland, McTavish and Pirro, 1973), and Textpack (cf. Tesch, 1990). At the same time, however, the technique of content analysis has come under criticism in some quarters for the decontextualization of words from the discourse being examined. Billig's (1989:206) criticism is typical: 'This sort of methodology can count words, but it cannot interpret them. Under some circumstances mere counting can lead to misleading conclusions.'

The need for the integration of content analysis with other approaches to text analysis in modern linguistics has been recognized for some time. Markoff, Shapiro and Weitman (1974:8), for example, observed that 'the linkages between content analysis and linguistics have been generally tenuous'. Content analysis needs to be incorporated into a broader approach to discourse analysis so that it may be seen to perform a clear role within such an approach rather than one which appears to be in competition, or is — as Markoff, Shapiro and Weitman (1974:7) put it — 'a methodological ghetto'. The level of vocabulary is clearly important in the analysis of discourse, but, as Billig's criticism suggests, words in discourse may only be interpreted precisely in the context in which they occur. A count of the word *good* in a text, for example, may be misleading: how many of these instances are negated and thus express the opposite of the concept 'good'; how many are discoursal interjections without any real content; moreover, to what or whom do the instances of *good* actually refer? What is needed, therefore, is a level of **relational** content analysis where the relationships between words can be defined and those relationships, in addition to the counts on individual words or categories, may

*This work was supported by a SERC/DTI grant number GR/F36385 IED4/1/1143 to G.N. Leech, J.A. Thomas and R.G. Garside. I wish to thank the grantholders, our commercial partners at Reflexions Market Research Ltd., and my colleague, Paul Rayson, for their several roles in this collaborative project. Further thanks are due to Geoff Leech for his helpful comments on the initial draft.

be classified and counted. This is what I hope to begin to describe in this paper.

Perhaps the earliest attempt to relate words within content analysis was the methodology of Evaluative Assertion Analysis (Osgood, 1959). However, Evaluative Assertion Analysis is only concerned with attitude direction towards certain entities and thus involves coding only a specific subset of relations, and moreover re-coding them in a particular way which is some distance removed from the actual text. But more basic syntactic relations have also been a long-term aim of computer-aided content analysis. The designers of the General Inquirer (Stone et al., 1966), for example, recognized the importance of syntactic relations in content analysis, but the level of parsing technology at that time meant that automated syntactic analysis was not possible, and thus the texts used by the program were annotated by hand to show syntactic relations. Despite early hopes, a parsing module was never in fact added, and indeed recent work on the incorporation of syntax into content analysis (e.g. Roberts, 1989; Lederer and Hudec, 1992) still largely employs manual annotation in this respect.

A recent attempt to define a computer-assisted link between content analysis and higher level textual representations has been that of Franzosi (1989) who suggests that researchers should adopt a new type of coding system based on the actual words used in the text, and which also links the actors with their respective actions. A formalism which incorporates this form of analysis, he argues, is the semantic text grammar. A text grammar is somewhat similar to a syntactic phrase structure grammar in its general form, but instead of using re-write rules to indicate the structure of syntactic constituents (e.g. noun phrases, verb phrases) it uses them to indicate the structure of conceptual units such as 'events' and 'actors'. According to Franzosi, text grammar analysis has the following advantages over traditional content analysis:

1. Text grammars are fuller and more explicitly relational than traditional content analysis coding.
2. They retain the story-like structure of the text and also the original lexis.
3. The codes are more reliable because of functionally-defined linguistic structures.
4. All data are recorded.
5. Analytic levels are clearly specified by the grammar.
6. The grammar is easy to implement computationally (especially in the form of a database).

Although some of Franzosi's criticisms of traditional content analysis are true (for example, the issue of compacting the content into a single count regardless of temporal situation or the actors involved has continued to prove problematic), there are alternative ways of alleviating some of these problems. There are in fact several disadvantages to the semantic grammar formalism as presented by Franzosi. Though the detail, retention of structure, retention of

original lexis and the coding of relations may be considered positive factors in increasing precision and reducing any perceived theory-dependent bias in content category systems, this also means that the quantitative aspect of content analysis is largely lost: Franzosi seems to be advocating a considerably less readily quantifiable approach to coding. Whereas Franzosi uses only a few limited codes in his grammar (e.g. actor, action), there are good reasons for retaining richer conceptual content classification above the level of the actual words used in texts. Apart from statistical arguments (it is very difficult to obtain valid results from standard statistical tests such as chi-square with the very small sample sizes which are quite likely to occur at the level of lexemes), people also tend to repeat the same concept within a discourse in somewhat different words through the use of virtual synonyms or the negation of a positive attribute, for example:

thoughtful	...	*not hasty*
lucrative	...	*well paid*
good	...	*brilliant*
very funny	...	*absolute riot*

(McCarthy, 1988:192).

This kind of repetition is central to obtaining the frequency counts of traditional categorial content analysis, but this conceptual similarity is lost in a purely word-based frequency count. It is also not easy to see how text grammar differs from traditional qualitative analysis and systematic abstracting, except in its strict formalism. All that it seems to offer is a way of indexing text in the form of a database with a pre-defined structure. Semantic text grammars are additionally anything but easy to implement computationally, except as fast annotation programs for human coders: because event packages are not tied to linguistic structures such as sentences or paragraphs, but may be distributed throughout the text, they are difficult to identify by anything other than human inspection.

II. Natural Language Processing Software

However, with the aid of modern linguistic software tools, it is possible to identify automatically certain relations within textual data. A considerable amount of research has been carried out in the field of corpus linguistics — that branch of linguistics which deals with the empirical study of large bodies of (usually machine readable) language — to develop software which will automatically annotate large bodies of textual data with various kinds of information.

A. Part-of-Speech Taggers

A number of packages have been produced to tag words in a text with the appropriate part of speech. Of particular importance is the package to which

I shall refer in the remainder of this paper, the CLAWS system. CLAWS (the Constituent Likelihood Automatic Word-tagging System) has been developed at the University of Lancaster's Unit for Computer Research on the English Language (UCREL) from the early 1980s onwards (Garside, Leech and Sampson, 1987). Although the structure of CLAWS has seen some changes since the first version was produced, it still consists of three stages: pre-edit, automatic tag assignment, and manual post-edit. In the pre-edit stage the machine-readable text is automatically converted to a suitable format for the tagging program. The text is then passed to the tagging program which assigns a part-of-speech tag to each word or word combination in the text using various heuristics to deal with words which are not in its dictionary. Because one orthographic form may have several possible parts-of-speech (e.g. *love* can be a verb or a noun), at this stage of the process CLAWS uses a probability matrix derived from large bodies of tagged and manually corrected texts to disambiguate the words in the text. The matrix specifies transition probabilities between adjacent tags, for example given that *x* is an adjective, what is the probability that the item to its immediate right is a noun? CLAWS tracks through each sentence in turn applying these probabilities. Finally manual post-editing may take place if desired to correct fully the machine output. The CLAWS system enjoys a success rate in the region of 96%-97% on written texts, and is also successful, though to a slightly lesser degree, on spoken texts. Success on spoken texts currently depends very much on their level of formality. 'Written to be spoken' texts perform rather better than impromptu conversation, primarily because part-of-speech transitions are more predictable, given that the current probability matrix is based on statistics extracted from written language: conversation tends to contain more 'fillers' (such as *um* and *ah*) and discontinuities where speakers suddenly break off and start afresh. A matrix for conversational speech has still to be constructed at the time of writing (April 1993).

B. Syntactic Parsers

Research has continued into the production of fully automatic parsers. A parser performs automatic syntactic (= grammatical) analysis of sentences in a text, i.e. it identifies noun phrases, verb phrases, etc. and how these relate to one another. An example of a corpus-based parsing system is the parser under development at Leeds University (Souter and O'Donoghue, 1991) which uses statistical probabilities, as the CLAWS tagging system does, in order to find a best fit analysis for a sentence. Parsing is important for the development of many kinds of natural language processing applications, but although statistically-based parsers such as the Leeds system are, on the whole, more successful than purely rule-based parsers, the parsing of natural corpus data as opposed to the invented examples of linguists is still an active research area. Parsers have a lesser success rate than part-of-speech taggers and thus at the present they generally do not form the basis of robust text processing systems. It is also questionable whether a full parse is necessary or economical for content analysis. The kind of relational analysis which I contend is needed for content analysis is both limited in scope and also

is partly reliant on semantic notions such as transferred negation[1] rather than on the traditional immediate constituent phrase structure grammars, or modern elaborations of such grammars, generally adopted in automatic parsers. The particular needs of content analysis and the current state of the art in parsing both suggest that a limited set of specific rules may be of immediately greater value than a full sentence parser.

III. The Lancaster Content Analyzer

A project at Lancaster University, ongoing since May 1990, has been aiming to develop an automatic content analyzer based on the CLAWS system and generally probabilistic methodologies (Wilson and Rayson, forthcoming). This system accepts as input language data in machine readable format (pure ASCII), which are then tagged for part of speech using the CLAWS part-of-speech tagging system. The tagged text is fed into the main semantic analysis program, which assigns semantic tags representing the general sense field of words from a lexicon of single words and an idiom list of multi-word combinations (e.g. *as a rule*). The tags for each entry in the lexicon and idiom list are arranged in general rank frequency order for the language. The text is manually pre-scanned to determine which semantic domains are dominant; the codes for these major domains are entered into a file called the 'disam' file and are promoted to maximum frequency in the tag lists for each word where present. This combination of general frequency data and promotion by domain, together with heuristics for identifying auxiliary verbs, considerably reduces mistagging of ambiguous words. After automatic tag assignment has been carried out, manual postediting takes place, if desired, to ensure that each word and idiom carries the correct semantic classification. A program then marks key lexical relations (see below), and a final module performs the automatic mapping of semantic tags into a set of content categories devised for a particular research project, and allows for statistical analysis and concordancing.

The assignment of content categories to individual words, together with statistical analysis, is as far as programs such as General Inquirer went. Indeed many content analysis systems apart from General Inquirer made little if no attempt at solving the very serious computational linguistic problems which content analysis highlights, for example word sense disambiguation. What I should like to argue for in this paper is the analytic importance of the linking of key relations which is performed in the Lancaster analyzer, which, together with the foundations of a more sophisticated approach to word sense disambiguation, makes it a considerable advance on programs such as General Inquirer.

[1]Transferred negation occurs when one of a particular set of verbs (e.g. *think*) introducing an indirect statement is structurally negated but semantically throws its negation forward to the main verb or adjective in the indirect statement; for example in the sentence *I do not think you notice it* the meaning is that I think you do not notice it, i.e. the act of noticing rather than thinking is the one which is negated (cf. Quirk et al., 1985:1033-1035).

IV. Key Relations

It is possible to see the analysis of textual data within social science as forming three levels, with the full textual detail accessible from each in an automatic analysis system by the use of a concordance/browser (figure 1). These

Figure 1: Levels of Textual Analysis

three levels are similar in many ways to the three levels which Fairclough (1989:110-111) identifies as important in critical language analysis (vocabulary, grammar, and textual structures). However, in this case, rather than identifying strictly linguistic levels, the three levels signify the varying number of linguistic levels employed in a particular kind of text analysis. At one extreme there is the possibility of carrying out a full critical discourse analysis in the sense of scholars such as Fairclough (1989) — this would involve a careful scrutiny of all the linguistic devices employed by the persons who produced the discourse, taking into account the social and ideological setting of the discourse. At the other is traditional content analysis, the counting of content categories or individual words, and making conclusions purely on sense frequency data. Discourse analysis is thorough and comprehensive, but it is very time consuming. It also requires specialist linguistic knowledge: the state of the art in automatic natural language processing is such that many of the needs of a full critical analysis — for example the relationships between sentences — cannot be satisfied by automatic means. Content analysis on the other hand is well established as a social research technique but it is a decontextualized method, that is, as Franzosi (1989) points out, it takes the linguistic content of texts at a high level of aggregation without consideration for the interrelationships between words or concepts. What I want to suggest is that the key to an integrated approach to textual data analysis is the middle level, that of identifying key relations.

By key relations I mean those interrelationships between words which are considered important in the interpretation of the discourse; in particular, these include the agents of actions in the text, the attributes assigned to various persons or things, and the various modifying and negating words and phrases associated with these.

The key relations with which the Lancaster system can currently cope are:

- **Nouns and adjectives**,
  e.g. AGE is IRRELEVANT; MEDICAL TECHNOLOGY
- **Adverbial modifiers and adjectives**,
  e.g. VERY NICE; QUITE GOOD
- **Negation**,
  e.g. they do NOT LIKE our values; payments are NOT GOOD

The system does not at present identify the subjects and objects of verbs (other than BE + adjective), or the referents of phoric pronouns such as *he*, *she* or *it*. Phoric relations will therefore made explicit at the stage where the text is converted to machine readable form, e.g. by using glosses in the international standard markup language SGML, thus:

```
Did you know
<annotation type=anaphor gloss="the mineral water">
it
</annotation>
was Swedish?
```

This may seem rather clumsy. However, given the frequency of assertions relating to phoric pronouns rather than explicit nouns, such referents need to be identifiable. The automatic identification of pronoun reference is beyond the current state of the art in natural language processing (although we are planning to work on this area) so these referents need to be inserted into text manually. Also it should be noted that the Lancaster system was designed with the analysis of in-depth survey interviews in mind, hence the annotation may be inserted at the stage of data transcription when the tape recordings are typed up into machine readable text, rather than as part of a postediting process on existing machine readable text. If the number of important referents is a small and largely closed set — perhaps because of interview domain — then a macro may be written in most word processors which will bind the insertion of all the SGML markup to a couple of key presses.

The linking of key relations in the text is performed on the basis of sequences of part-of-speech tags assigned by the CLAWS tagging suite in the first phase of the data processing. It will be recalled that CLAWS assigns a unique part-of-speech tag to each word or syntactic idiom in a text. Through a careful analysis of linear sequences of such tags in sample texts (a corpus of approx. 79,000 words of market research interviews), it was possible to determine for each type of key relation the nature and frequency of the characteristic sequences of part-of-speech tags within which these relations occur, then to specify the links on the basis of these sequences. The system thus does not attempt to parse phrases: it seeks merely to link two key items.

My colleague, Paul Rayson, wrote a program which can implement these rules on the tagged text by reading them in from a rule file into which they have been entered using a specific formalism.

There are two kinds of rule used to link the key items:

A. Match and Link

Match and link rules attempt to match a fixed template of part-of-speech tags within the sentence. These may include 'wildcards' — markers which will match one or more characters — to allow for paradigmatic variation in a part of speech, (e.g. the base form, comparative and superlative of adjectives), or the interpolation of any word or words. The rules also allow for the specification of occasionally intervening parts-of-speech (by the use of bracketed expressions), and multiple strings of certain parts-of-speech, indicated by *n* (= any number of). For example, the following rule is a match and link rule[2]:

```
NOT[.]  (ADVERB n)  VERB[.]
```

The full stop in square brackets indicates the words which are to be linked. This rule will link *not/n't* negatives with their verbs, including links across any number of possible adverbs, in phrases such as:

Family values alone will NOT FEED a hungry child.
He can NOT really CHANGE America.
She has NOT very often GONE there.

B. Search and Link

A search and link rule, rather than matching an invariable template frame, performs a dynamic search within the sentence for one of the two items it is to link to. These searches are indicated by a statement in square brackets on the member of the linked pair whose position in the part-of-speech sequence is fixed, in place of the full stop in the match and link rules. This consists of an < or > sign indicating a backward or forward search from that point, and the tag, or range of tags, which is to be searched for (listed in preference order, i.e. the second tag in the list is only searched for if there is no match for the first, and so on). Thus for example:

```
PART-OF-"TO BE" (ADVERB n) (NOT) (ADVERB n)
ADJECTIVE[<PROFORM/DETERMINER/NOUN]
```

will match adjective collocations with nouns / pronominals / determiners such as:

CHANGE is CERTAIN.
GOVERNMENT is too BIG.
THIS is not UNUSUAL.
IT is not very TASTY.
SOME are GOOD

Both match and link and search and link rules may also indicate specific words in the rule schema in place of the broader part-of-speech declarations:

[2]In the examples in this paper, for ease of comprehension I have used words rather than the actual CLAWS tags which are used in the program's rule file.

this is particularly important in transferred negation (cf. Wilson, 1991) and in *no* and *never* negation (where *no* and *never* do not receive the same negative tag as *not/n't* in the CLAWS tag set).

The key relation linkage rules in the Lancaster system exhibit a high degree of success, within the 90s per cent. Most of the errors are in fact caused by part-of-speech mistagging by CLAWS, given that the current probability matrix and idiom list is designed for written rather than conversational language, and not by failures of the linkage rules themselves. The rules perform even better on texts which have been manually postedited. This is a highly satisfactory success rate for the fully automated processing of unrestricted natural language. Nevertheless, these are a limited set of relations extracted from a relatively grammatically simple genre of English (unprepared colloquial conversation). Certain problems may be foreseen in attempting to apply or extend these rules on more grammatically complex genres. For example, it is particularly important to identify the actors in the discourse and the people/objects to which attributive statements refer. Linguistically, these are encoded as nouns or pronominals. The main problem for linkage is that nouns may be modified by phrases including other nouns, for example in the following sentence the noun *colours* is modified by a phrase containing the nouns *cereal* and *packet*:

The colours on the cereal packet are bright.

This means that a simple template or unconstrained 'search for noun' rule will result in only part of a noun phrase being linked, and occasionally this may be the wrong part, i.e. not the head noun. These errors are not problematic in the sense that the head may be recoverable from examining the context of the other, incorrectly linked noun in a concordance/browser, but they are serious if one wants to perform automatic statistical counts at the level of key relations, either on words or on content categories. It is therefore imperative that rules are devised which will analyze, at least in a limited way, the internal structure of a noun phrase. Tentative research suggests that relatively successful probabilistic rules might be extracted in a similar way to the linkage rules, although this aspect requires much further work, and a combination of approaches, possibly including semantic valency information (e.g. Leech, 1986), might prove more successful for this aspect of a linkage program.

When linkage rules have been applied, one has a text which is annotated with a number of word pairs as in the following, the links being indicated by numbered # signs:

```
She said that the railway#1 was not#2 very#3 efficient#1#2#3.
```

What these word pairs result in is a network of links, which may perhaps be visualized as labelled arcs. Some words will only have one end of an arc attached to them, whereas others may have two, three or more (as with *efficient* in the example above). Where a word has links to other words, these links need to be extracted from the text into a collocation of the relevant words. Where multiple links occur on a word, the words at the other end of those arcs should all be included in a single collocation with the multiply linked word. These various collocations of attribute assignments can then be extracted at a data

inquiry stage, and quantified if required either at the level of the individual words or at the level of content categories; for example, one could ask how many times content category 12 refers to *Lake District* in a body of text, or, if desired, more specifically how many times a particular word within that category refers to *Lake District*, then break these counts down according to the various collocations of modifiers and negators.

V. Conclusion

I have demonstrated in this paper a successful technique for the linkage of key relations in text based on a state of the art part-of-speech tagging system. This technique forms an important extension of the method of content analysis and represents a bridge by which content analysis may be more closely integrated with discourse analysis in the social scientific analysis of textual data. Not only does it permit the interrelationships of words or content categories to be defined in a way which has previously been possible only by manual analysis of large quantities of text, it also provides a means by which a further stage of critical discourse analysis — the linking of nouns with their attributes — may be automated and quantified. Indeed, it is likely that linkage rules may be devised to handle further key relationships in discourse. The linkage approach promises to be an important addition to the tools available to social scientists needing to analyze large quantities of text, and a key in finally bringing the rival camps of content and discourse analysis together. It may perhaps also be influential in beginning to establish a more quantitative approach to discourse analysis than has hitherto been feasible. Automation enables the human analyst to examine more data in more detail than before, and to do so in a way strictly answerable to objective evidence. Such integration of techniques will be an important step forward for sociological methodology, and if these ends can be achieved, then this research will have been more than worthwhile.

References

Billig, M. (1988). Methodology and Scholarship in Understanding Ideological Explanation. In: C. Antaki (ed.), *Analysing Everyday Explanation: A Casebook of Methods*. London: Sage.

Cleveland, C.E., D. McTavish and E. Pirro. (1974). Quester: Contextual Content Analysis. Paper presented at ISSC/CISS Workshop on Content Analysis in the Social Sciences, Pisa, September 1974.

Fairclough, N. (1989). *Language and Power*. London: Longman.

Franzosi, R. (1989). From Words to Numbers: A Generalized and Linguistics-based Coding Procedure for Collecting Textual Data. *Sociological Methodology* 19, 263-298. Oxford: Blackwell.

Garside, R., G. Leech and G. Sampson (eds). (1987). *The Computational Analysis of English: A Corpus Based Approach*. London: Longman.

Iker, H.P. and N.I. Harway. (1969). A Computer Systems Approach toward the Analysis of Content. In: G. Gerbner et al. (eds), *The Analysis of Communication Content*. New York: John Wiley.

Lederer, B. and M. Hudec. (1992). *Computergestützte Inhaltsanalyse*. Frankfurt a.M.: Campus Verlag.

Leech, G.N. (1986). Semantico-Syntactic Analysis of a Corpus for Speech Recognition. Unpublished manuscript.

Markoff, J., G. Shapiro and S.R. Weitman. (1974). Toward the Integration of Content Analysis and General Methodology. *Sociological Methodology* 1975. San Francisco: Jossey-Bass.

McCarthy, M. (1988). Some Vocabulary Patterns in Conversation. In: R. Carter and M. McCarthy (eds), *Vocabulary and Language Teaching*. London: Longman.

Osgood, C.E. (1959). The Representational Model and Relevant Research Methods. In: I. de S. Pool (ed.), *Trends in Content Analysis*. Urbana, IL: University of Illinois Press.

Quirk, R. et al. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Roberts, C.W. (1989). Other than Counting Words: A Linguistic Approach to Content Analysis. *Social Forces* 68:1, 147-177.

Souter, C. and T.F. O'Donoghue. (1991). Probabilistic Parsing in the COMMUNAL Project. In: S. Johansson and A.-B. Stenström (eds), *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter.

Stone, P.J. et al. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Tesch, R. (1990). *Qualitative Research: Analysis Types & Software Tools*. London: The Falmer Press.

Wilson, A. (1991). 'No', 'Not' and 'Never' Negation in a Corpus of Spoken Interview Transcripts. *Lancaster Papers in Linguistics*, number 73.

Wilson, A. and P. Rayson. (forthcoming). Automatic Content Analysis of Spoken Discourse. To appear in: C. Souter and E. Atwell (eds), *Corpus Based Computational Linguistics*. Amsterdam: Rodopi.