Analysing Discourse Topics and Topic Keywords

Abstract

Discourse topic is an intractable and inherently subjective notion making analysis problematic. This paper overcomes some of the problems by treating topic as a fuzzy concept and views discourse topics as sets of topic keywords. The study examines the identification of topic boundaries and topic keywords by informants and by four methods of analysing topics – topical structure analysis, given-new progression, lexical analysis, and topic-based analysis. Comparing the findings from these four methods against those from the informants, it was found that given-new progression is the most valid method for identifying topic boundaries, and topic-based analysis is the most valid for identifying topic keywords. There are also notable differences in the types of keywords identified and the bases for identifying keywords between the methods and the informants.

1. The nature of discourse topics

Discourse topic is one of the most intractable notions in linguistics. Most previous studies have either treated discourse topic as a pretheoretical notion (e.g. Brown and Yule 1983b) or relied solely on the researcher's intuitions in dealing with topics (e.g. Maynard 1980; Carlson 1983; Crow 1983; Rost 1994; Shepherd 1998). Given the potential importance of the notion in many areas of linguistics, the lack of well-founded methods of addressing discourse topics is surprising.

Topic is a term used with several meanings in linguistics. The main differences in meaning relate to the level of discourse to which *topic* is applied. At the most local level, sentences have topics in a topic-comment approach (Chafe 1976; Jäger 2001), analogous to theme-rheme analyses (see Section 1.1). Also at a local level, some researchers working with centering theory (e.g. Walker et al. 1998) have argued that backward-looking centers are local-level topics (but cf. Hu and Pan 2001). At the other extreme, we could talk colloquially of the topic of a book or a lecture. In this paper, I will use *discourse topic* (hereafter *topic*) at an intermediate level to refer to the topics of ideationally coherent stretches of discourse ranging from a single sentence to a couple of paragraphs.

At this mid-level, the coherence of topics can be seen as deriving from a set of propositions delimiting a certain range of semantic space (van Dijk 1977; Crookes and Rulon 1988), or from "a clustering of concepts which are associated or related from the perspective of the interlocutors in such a way as to create relevance and coherence" (Watson Todd 2003: 22). While primarily conceptual in nature, it seems likely that topics will be indicated by some linguistic, probably semantic, features in a text. The extent to which linguistic features are used by people to identify topics, however, is unclear.

The propositional and conceptual bases for identifying topics can lead to differences in how topics can be expressed. With a propositional approach, a topic would be expressed as a proposition (see also Keenan and Schieffelin 1976), whereas a conceptual approach allows topics to be expressed as noun phrases. Two other ways of expressing topics have also been suggested: McNally (1998) argues that topics should be expressed as questions, and the intimate bond between lexical items and topics (McCarthy 1991) suggests that key lexical items can be indicative of topics. In this paper, topics will be expressed in two ways: as noun phrases since most previous studies of topic (e.g. Coulthard 1977; Maynard 1980; Crow 1983; Gardner 1987; Mäkinen 1992; Rost 1994) have done this, and as sets of keywords since this allows different analyses of topics to be compared easily.

In addition to identifying and expressing the topics of discourse segments, a second focus of previous research into topics has concerned identifying the lexical and phonological indicators of boundaries between topics (e.g. Covelli and Murray 1980; Stech 1982; Richards and Schmidt 1983; Hemphill 1989). Such indicators, however, only identify the clear-cut boundaries associated with abrupt topic shift. Less clear is how to identify boundaries when one topic gradually melds into another, so-called topic drift (Crow 1983).

This paper takes topics to be related clusters of concepts expressible either as noun phrases or as sets of keywords, and examines both the ways in which topics of discourse segments can be identified and how boundaries between topic segments for both topic shift and topic drift can be identified.

1.1 Methods of analysing topics

Most previous work directly addressing discourse topics was conducted in the 1970s and 1980s as part of the text linguistics movement. This work has continued within natural language processing (e.g. Gardent and Webber 1998; Ferret and Grau 2000), but the need to make any analyses computerized limits its applicability. Within mainstream linguistics, on the other hand, surprisingly little has been done. However, several well-known approaches within linguistics do have implications for topics and, with minor adjustments, can be used to identify the topics of discourse segments and boundaries between topics.

The first approach is related to sentence-level topics in a topic-comment approach. Theme-rheme or topical structure analysis similarly divides sentences into two parts, the theme and the rheme, where the theme is "what the sentence is about" and the rheme is "what is said about [the theme]" (Connor 1996: 81). There are various criteria for identifying themes in English (e.g. Brown and Yule 1983a; Fries 1983; Davies 1994; McCarthy and Carter 1994), and, in this study, I will follow Halliday's (1967) influential criteria for identifying themes as "clause initial elements up to and including the first ideational element" (Berber Sardinha 1997: 69). Having identified themes and rhemes in sentences, we can examine how the themes and rhemes of succeeding sentences are related, since themes provide an organization for the discourse with rhemes providing the message that pushes the communication forward (Daneš 1974). There are various permutations of theme-rheme progression, such as parallel progression where succeeding sentences have the same theme (Lautamatti 1978; Connor and Farmer 1990; Schneider and Connor 1990). Where no theme-rheme progression is apparent, a coherence break (Wikborg 1990) indicative of a boundary between topics can be identified. Examining the themes between coherence breaks can lead to the identification of the topic for that stretch of discourse (Watson Todd 2003).

While theme-rheme analyses are primarily linguistic, a similar approach dividing sentences into two parts involving both psychological and linguistic concerns is given-new progression. In this study I will follow Chafe's (1976, 1980) definitions of the given-new distinction where given information is information in the consciousness of the recipient and new information is information being introduced into the recipient's consciousness, since this approach has implications for topics in that such a given-new distinction is related to the "current discourse space" (Langacker 1996: 334). Although it is impossible to definitively identify information in the recipient's consciousness, certain characteristics of language can help us to identify given and new information. For example, ellipted material (Chafe 1980; Tomlin et al. 1997), pronominalized material (Chafe 1976; Clancy 1980; Palmer 1981), and noun phrases with definite articles (Haviland and Clark 1974) are all given information. After identifying given and new information, a similar approach to theme-rheme progression involving given-new progression (Goldberg 1983; Firbas 1987; Rutherford 1987) can be used, so that boundaries between topics are shown by given-new coherence breaks and the topics of stretches of discourse between these boundaries are indicated by given information (Watson Todd 2003).

A different approach to identifying topics involves creating networks of lexical items in a discourse with density of linkage in the networks being indicative of topics (de Beaugrande and Dressler 1981). An example of this approach is Hoey's (1991) lexical analysis which examines how recurrences of lexical items across sentences reflects text organization. Repetitions or paraphrases of lexical items provide cohesive links between sentences, and two sentences with the number of links above a certain predetermined level are termed bonded sentences. Hoey argues that sentences with high numbers of bonds with subsequent sentences but no preceding sentences are topic-opening (and vice versa for topic-closing sentences). From this, topic boundaries can be identified before sentences which do not link to nets of bonds created by previous sentences, and topics can be identified from weightings of the lexical items providing links within a cluster of bonds.

A further approach relies on hierarchies, rather than networks, to represent discourse. Topic-based analysis (Watson Todd 1998) involves identifying the key concepts in a discourse based on frequency (see Scott 1997) and then drawing up a hierarchy showing the relationships between these concepts using loose interpretations of semantic relations such as hyponymy and meronymy. The sequence in which the concepts appear in the discourse is then mapped onto this hierarchy. Moves between concepts are assigned a distance in semantic space based on the distance between the two concepts in the hierarchy. Large distances are indicative of coherence breaks or topic boundaries, and the key concepts between boundaries can be weighted to identify topics.

2. Purposes of the study

There are, then, several methods in linguistics that can be used to identify topics and topic boundaries, and this study examines four approaches: topical structure analysis (TSA), given-new progression (GNP), the lexical analysis of Hoey (LA), and topic-based analysis (TBA). To investigate the extent to which linguistic features influence people's identification of topics, findings concerning the location of topic boundaries and the identification of topics from each approach are compared with the topics and topic boundaries identified by human informants for the same discourse.

This study therefore aims to do the following:

- To compare the identification of topics and topic boundaries by different informants;
- To compare the identification of topics and topic boundaries by four methods of analysis;
- To compare the identification of topics and topic boundaries between informants and methods of analysis.

These comparisons should provide insights into how the informants identify topics. If there is a close match between the informants and one particular method, it seems likely that the textual features analysed by that method are of greater import in identifying topics than the textual features of the other methods. Given that the informants' identification of topics involves an interaction between the informant and the text whereas the four methods of analysis rely on surface linguistic features of the text, the extent of any matches and the amounts of variation should also shed light on the extent to which linguistic features are used in identifying topics.

3. Methodology

3.1 The text

To investigate topics and topic boundaries, we need a text which prioritizes the ideational metafunction. In order to elicit informants' identifications of topics and topic boundaries, this text should be a written text. However, the original work in two of the four methods (GNP and TBA) was conducted on spoken discourse suggesting that a text with the characteristics of spoken language may be easier to analyse. Therefore, a transcript of an excerpt from the film *An Inconvenient Truth* in which Al Gore presents arguments about global warming was used. Since this transcript is available as a written text on a website, it can be treated as a written text while still retaining many of the characteristics of spoken language.

The chosen excerpt was divided into T-units (see Fries 1994) rather than sentences since the identification of sentence boundaries in transcribing a spoken text is somewhat arbitrary. The excerpt is 175 T-units long. In order to be able to apply given-new progression and lexical analysis, referents for referring expressions and ellipted material need to be recovered and this was done following the guidelines of Watson Todd (2003).

3.2 The informants

To gain insights into the identification of topics and topic boundaries, 7 educated native-speaker informants were asked to identify topics for the text. The potential problems of defining mid-level topics in a comprehensible way for the informants were avoided by describing discourse topics as "something between the topic of a sentence and the gist of a text" identifiable most usually for a stretch of several sentences and "most usually identified as a noun phrase". To make this concrete for the informants, an example of another excerpt from *An Inconvenient Truth* was given in two columns, where the first column contained the text divided into T-units and the second column contained intuitively assigned topics for this text identified by the researcher. Table 1 contains an extract from this model. The 175-T-unit text used as data in this study was then shown with a column for the informants to fill in the topics they identify.

Table 1 Model extract for topic identification

Text	Topic
1. This brings me to the second canary in the coal mine, Antarctica, the largest mass of ice on the planet by far.	Antarctic ice shelves
2. A friend of mine said in 1978, "If you see the break up of ice shelves along the Antarctic Peninsula, watch out, because that should be seen as an alarm bell for global warming."	
3. If you look at the peninsula up close, every place where you see one of these green blotches is an ice shelf larger than the state of Rhode Island that has broken up in just the last 15 to 20 years.	
4. I want to focus on just one of them called Larsen B.	Larsen B
5. I want you to look at these black pools here.	Black pools
6. It makes it seem almost as if we are looking through the ice to the ocean beneath.	
7. But that's an illusion.	
8. This is melting water that forms this pool.	

Note: the full extract used as a model is 34 T-units long.

3.3 Data analysis

For the informants and methods of analysis, boundaries between topics were identified as follows:

- For the informants, boundaries were identified at the end of a T-unit before a new topic.

- For topical structure analysis, boundaries were identified at coherence breaks where no theme-rheme progression was apparent.

- For given-new progression, boundaries were identified at given-new coherence breaks where no given-new progression was apparent.
- For lexical analysis, the number of links needed to create a bond was set at 2, and boundaries were identified at points in the network where there were no bonds with the preceding 10 T-units and no links with the preceding 2 T-units.
- For topic-based analysis, boundaries were identified at points where a move had a distance of more than 2 in the hierarchy of key concepts.

In identifying boundaries in these ways, a certain amount of arbitrariness enters the analysis, especially for lexical analysis and topic-based analysis. While it is possible to set both number of links to create a bond and distances in the hierarchy at 1, 3 or 4, the resulting lengths of monotopical stretches of discourse would be of a different order of magnitude to those identified by the informants and the other two methods of analysis. The cutoff points for identifying topic boundaries were therefore set to ensure comparability between different methods and between methods and informants. The frequency and locations of topic boundaries were compared between informants, between methods of analysis, and between informants and methods of analysis.

Topics are identified by the informants as noun phrases, but by the methods of analysis as sets of keywords. The noun phrases identified by the informants can be analysed for exact matches, but further analyses require the noun phrases to be converted into sets of keywords. To do this, function words were ignored, and the remaining words given weightings to a total of 10 with priority given to content words over shell nouns indicating macro-functions (see Aktas and Cortes, 2008) and to words appearing frequently in the stretch of discourse. Similar weightings were given to the keywords identified by the methods of analysis. In this study, topics are taken as being subjective with no definitive identification of a single topic of a stretch of discourse being possible. Rather, a fuzzy logic approach (see Watson Todd 2005) is taken with the weightings of keywords being indicative of the likelihood of their being identified as aspects of the topic. The more heavily weighted a keyword, the more likely that it should be considered an aspect of the topic for that stretch of discourse. The weightings of the various keywords identified by all the informants can be totaled to give estimates of the probabilities of the keywords being aspects of the topic for each T-unit. These total weightings can be compared against the keyword weightings for each method of analysis and for all methods combined. Even if topics are regarded as subjective, for some stretches of discourse there may be very high levels of agreement about the topic, while for other stretches several competing topics could be possible. The former can be identified as T-units where one keyword is far more heavily weighted than others, and the latter as T-units where several keywords have similar weightings.

This can be measured from a fuzzy logic perspective using fuzzy entropy or (A intersect not-A)/(A union not-A) (see Kosko 1993; Watson Todd 2005) which shows the amount of uncertainty of a fuzzy set. The fuzzy entropy of the topics of each T-unit was calculated for both the informants and the methods of analysis and compared with low values indicating high levels of agreement on a topic. Finally, the types and variety of keywords identified by the informants and by the methods were compared.

4. Findings

4.1 Topic boundaries

The informants identified different numbers of topic boundaries in the extract as shown in Table 2. The average number of boundaries per informant is 31.57, and the average number of boundaries per informant per T-unit is 0.18.

Table 2 Number of topic boundaries identified by informants

Informant	*No. of boundaries*
A	19
B	34
C	41
D	38
E	31
F	23
G	39
Total	225

Even though the informants identified different numbers of boundaries, there is some agreement about where these boundaries are located in the text (see Table 3). For 13 points in the text, at least 5 of the 7 informants identified a boundary (a level of agreement vastly higher than would occur by chance), suggesting a high likelihood of a boundary really existing at those points. On the other hand, there are 90 points in the text at which it is very unlikely for there to be a boundary.

Table 3 Number of informants agreeing on locations of topic boundaries

No. of informants agreeing	No. of points in discourse
7	3
6	5
5	5
4	10
3	15
2	17
1	30
0	90

In addition to identifying points in the text where changes in topic occur, the information on topic boundaries also allows us to see the amount of variation in the lengths of monotopical stretches of discourse between boundaries by comparing the mean and standard deviations of these stretches. From Table 4, we can see that all informants had similar high levels of variation in the lengths of monotopical stretches of discourse.

Table 4 Variation in length of stretches of discourse between topic boundaries by informant

Informant	*Mean length of stretch*	*SD of stretch length*	*SD/mean of stretch length*
A	8.75	5.20	1.68
B	5.00	3.28	1.52
C	4.17	2.91	1.43
D	4.49	2.76	1.63
E	5.47	2.78	1.97
F	7.29	6.07	1.20
G	4.37	2.60	1.68

The findings concerning topic boundaries from the methods of analysis are similar. The average number of topic boundaries is 27.25, and the average number of boundaries per method per T-unit is 0.16 (see Table 5), both comparable to the averages for informants.

Table 5 Number of topic boundaries identified by methods of analysis

Method	*No. of boundaries*
TSA	23
GNP	31
LA	27
TBA	28
Total	109

Similarly, there are 14 points in the text where at least 3 of the 4 methods identify a boundary (see Table 6), and 114 points at which it is very unlikely for there to be a boundary – again, figures comparable to those of the informants.

Table 6 Number of methods of analysis agreeing on locations of topic boundaries

No. of methods agreeing	No. of points in discourse
4	5
3	9
2	15
1	32
0	114

For variation in the lengths of monotopical stretches of discourse between boundaries, however, although there is little difference between the methods (see Table 7), the figures comparing the mean with the standard deviation are generally lower than for the informants, suggesting that there is less variation in the lengths of stretches of monotopical discourse for the methods than for the informants.

Table 7 Variation in length of stretches of discourse between topic boundaries by method of analysis

Method	*Mean length of stretch*	*SD of stretch length*	*SD/mean of stretch length*
TSA	7.29	8.44	0.86
GNP	5.47	5.62	0.97
LA	6.25	5.34	1.17
TBA	6.18	4.90	1.26

To compare the findings concerning topic boundaries between the informants and the methods of analysis, we can look at how all methods together compare and how each method compares individually. For all methods, each point in the discourse where a boundary could occur can be rated from 0 to 7 for the informants depending on how many informants identify a boundary at that point, and from 0 to 4 for the methods. Comparing these two sets of figures for the whole text using the correlation coefficient, we find $r = 0.32$ ($N = 174$; $p < 0.01$) suggesting a significant, if not particularly high, overall level of agreement concerning the location of topic boundaries.

For each individual method, the ratings from 0 to 7 for the informants can be compared against whether each method identifies a boundary or not using point biserial correlation. The results are shown in Table 8 and, if we take the informants' locations of boundaries as a benchmark, suggest that given-new progression is the most valid method for identifying boundaries.

Table 8 Number of boundaries identified after each T-unit: correlations for total for informants against each method

Method of analysis	r_{pbi}	t	p
TSA	0.159	2.11	p<0.05
GNP	0.307	4.23	p<0.01
LA	0.129	1.71	p<0.05
TBA	0.152	2.02	p<0.05

Focusing only on those points where a given method identifies a boundary, we can look at the number of informants who agree with the location of a boundary at each point (see Table 9). Again, given-new progression shows the greatest match with the informants' identification of boundaries.

Table 9 Average number of informants agreeing with locations of topic boundaries identified by methods

Method	Mean no. of informants agreeing
TSA	2.00
GNP	2.45
LA	1.81
TBA	1.89

Overall, there appear to be reasonable levels of agreement in the overall frequency of topic boundaries and their locations between informants, between methods, and between informants and methods with given-new progression being the most valid of the methods of analysis concerning topic boundaries. However, there is more variation in the lengths of stretches of monotopical discourse for informants than for methods.

4.2 Topics

All of the topics identified by the informants were expressed as short noun phrases. Comparing these noun phrases is problematic. For example, for the last 4 T-units of the text, 2 informants identified the topic as *moral issue*, presumably equivalent to another informant's identification as *a moral issue*. One further informant identified the topic as *a moral issue not a political one*, for which it is less clear whether it should be considered equivalent. To enable more certain comparisons between topics, the keywords in the topics were identified and weighted to give a total of 10. Thus, for both *moral issue* and *a moral issue*, there are two keywords, *moral* and *issue*, each weighted 5; and for *a moral issue not a political one*, there are two keywords, *moral* and *political*, weighted 3 and one, *issue*, weighted 4 since this is a noun and appears twice in the topic. Converting topics into sets of topic keywords allows comparison between informants. Looking at the most heavily weighted keyword(s) for each T-unit, we find that there are reasonable levels of agreement between informants regarding the most important keyword (see Table 10). Indeed, there are no T-units for which at least two of the informants do not agree on the most important keyword.

Table 10 Number of informants agreeing on most heavily weighted keyword per T-unit

Max. no. of informants agreeing	No. of T-units	% of total T-units
7	25	14.29
6	14	8.00
5	27	15.43
4	59	33.71
3	46	26.29
2	4	2.29
1	0	0.00

For the methods of analysis, topics are identified as sets of topic keywords (so there is no need to convert noun phrases into sets of keywords). There appears to be slightly less agreement concerning topic keywords between methods than between informants (see Table 11). For the majority of T-units, the four methods of analysis do not fully agree on all of the keywords, and for about half of the T-units, only 2 methods agree on the most important keyword.

Table 11 Number of methods agreeing on keywords per T-unit

No. of methods agreeing	No. of T-units with agreement for all keywords	No. of T-units with agreement for most important keyword
4	0	29
3	12	49
2	41	87
1	122	10

By taking topics as sets of topic keywords, we can calculate total weightings for each T-unit for all informants and for all methods of analysis, allowing comparison between the two. From this we find that the informants and the methods agree on the most heavily weighted keyword for 37 T-units (or 21.14% of all T-units). Comparing the total weightings for all keywords for each T-unit between the informants and the methods, we find significant, if not high, levels of agreement (N = 2160; $r = 0.29$; $p < 0.01$). For the cumulative weightings of keywords across the whole text, levels of agreement are higher (N = 279; $r = 0.63$; $p < 0.01$). We can also compare the weightings of keywords for each method against the weightings for all informants for each T-unit (see Table 12). Topic-based analysis shows the greatest levels of agreement with the informants, with lexical analysis also agreeing fairly strongly. Topic keywords identified through topical structure analysis and given-new progression, on the other hand, have little relationship with those identified by the informants.

Table 12 Correlations for identifying keywords between informants and each method

Method	N	*r*	*p*
TSA	1899	0.07	n.s.
GNP	1921	0.13	n.s.
LA	1833	0.26	$p < 0.01$
TBA	1809	0.31	$p < 0.01$

A final issue concerning the topics identified for each T-unit concerns the extent to which informants or methods agree on a topic for a stretch of discourse. The probability of a definitive topic being identified for a given T-unit is measured through fuzzy entropy (following Watson Todd 2005) where 0 indicates complete agreement on a topic and 1 indicates an equal possibility of two or more potential topics being identified. For the informants, the mean fuzzy entropy per T-unit is 0.57 with a minimum of 0.12 and a maximum of 1.00, and for the methods the mean fuzzy entropy per T-unit is 0.59 with a minimum of 0.08 and a maximum of 1.00. Despite these overall similarities, there is very little agreement between the informants and the methods on the level of fuzzy entropy of each T-unit (*r* = 0.09; not significant).

Overall, there is some agreement among informants and among methods of analysis concerning the topic of each T-unit (expressed as sets of weighted topic keywords). There is also some agreement between the informants and each method with topic-based analysis and lexical analysis showing fairly high agreement. However, there is no agreement between the informants and the methods concerning the extent to which a definitive topic can be identified for each T-unit.

4.3 The nature of topic keywords

The total number of topic keywords for all T-units is very different for the informants and for the methods of analysis (even taking into account that there are more informants than methods). The informants identified a total of 145 different keywords (with complex repetitions (Hoey 1991) being counted as one keyword), and the methods of analysis identified 50 different keywords. In total, there are 239 different content words, or potential keywords, in the text.

While all 50 keywords identified by the methods appear in the text (as would be expected given the ways in which keywords are identified by the methods), 40 of the 145 different keywords identified by the informants do not appear in the text. These 40 keywords appear to fall into one of two categories. First, some keywords identified by the informants are paraphrases of concepts in the text. For instance, one informant identified *drift* (as in continental drift) as a keyword for a stretch of discourse which included *They [continents] moved apart from one another, but at one time they did in fact fit together*, even though the word *drift* does not appear in the text. Second, several keywords identified by the informants, such as *restatement*, *theory* and *description*, are shell nouns showing the rhetorical purpose of a stretch of discourse rather than its content. Of the 20 rhetorical words identified as topic keywords by the informants, only 6 appear in the text, and only one is identified as a keyword by the methods.

A further difference between the informants and the methods involves the emphasis placed on people's names as keywords. Both identify three names (including *Gore* for the first person singular) as keywords, but the weightings for these names are very different (see Table 13) with the methods of analysis placing a far greater emphasis on people's names than the informants. If we discount people's names and rhetorical words from the analysis, we find that the most heavily weighted keywords identified by the informants and by the methods agrees in 79 T-units or 45.14% of the text (compared with 37 T-units if they are not discounted). Similarly, if we consider only content words (and ignore both people's names and rhetorical words), the level of agreement on the overall weightings of keywords across the whole text is very high (N = 256; *r* = 0.91, compared with N = 279; *r* = 0.63 when all keywords are considered).

Table 13 Numbers and weights of word types

Word type	Informants		Methods	
	No. of words	Total weight	No. of words	Total weight
Content	122	9069.38	46	6931.43
Person	3	337.14	3	3054.29
Rhetorical	20	593.47	1	14.29

The identification of words not in the text as topic keywords by the informants implies that the informants are not basing their identification of topics solely on the frequency of content words in the text. This is confirmed when we compare the frequency of all content words in the text with the overall weightings of the keywords identified by the informants (N = 239; *r* = 0.65). In contrast, the overall weightings of the keywords identified by the methods has a very high level of agreement with the text frequency of the content words (N = 239; *r* = 0.87). The difference between these two correlations suggests that the informants rely less on frequency as the basis for identifying keywords than the methods do. Although the methods do take some account of saliency (superordinate keywords in the topic hierarchy in TBA, words in T-unit themes in TSA, and words as given information in GNP), the main method of identifying keywords in the methods is frequency. For the informants, on the other hand, there are instances where saliency clearly takes precedence over frequency. In the last 4 T-units, 6 of the 7 informants identified *moral* and *issue* as heavily weighted keywords even though they appear in that stretch of discourse (and, indeed, the whole text) only once, whereas *Congress* appears twice. There are, then, clear differences in the numbers of keywords, their nature, and the basis for identifying them between the informants and the methods of analysis.

5. Discussion

In this paper *topic* is viewed as a fuzzy concept in a way similar to how Hoey (1991) views coherence. He states that coherence is "subjective and judgments concerning it may vary from reader to reader" (p. 12), but also that "an overwhelming consensus" (p. 266) of opinion can be achieved. Taking such a view precludes the definitive identification of topics. Rather, we are interested in the extent of agreement and disagreement on topics by informants (and methods of analysis), and the findings indicate the likelihood of placement of topic boundaries and identification of topics. Comparing the findings from the methods of analysis to those of the informants can shed light on how the informants identify topics. If the likelihoods concerning topics identified by the informants differ greatly from those identified by a method of analysis, it would appear that the method of analysis is identifying something different to what people identify as topics.

Generally, the levels of agreement among informants concerning both the placement of topic boundaries and the most important keywords for T-units suggest that a fuzzy approach to topics is valid. Overall, although the amount of agreement between informants is much greater than chance would allow, there are very few points in the text where all informants agree. This suggests that the text may provide some common ground, but that individual informants' interpretations are different.

textual influences on topic identification are also shown by some noticeable agreements between the informants and the methods for placement of topic boundaries (especially for given-new progression) and for weightings of keywords (especially for topic-based analysis). However, the keywords identified by both topical structure analysis and given-new progression do not significantly agree with those identified by the informants. These findings have potential implications for other research into topics and coherence. The most commonly used research method for identifying topics and coherence is topical structure analysis (e.g. Witte 1983; Connor and Farmer 1990; Lee 2002). For instance, Lee examined the effects of teaching coherence to second language writers by measuring the coherence of their writing using topical structure analysis (even though the teaching had concerned given-new progression). The only previous research I am aware of comparing the effectiveness of different methods of analysing topics and coherence (Watson Todd 2003) found that topical structure analysis is the least effective method. The findings from the present study suggest that, instead of using topical structure analysis, if topic boundaries or coherence breaks are a concern, given-new progression should be used, and, if the actual topics are the focus, topic-based analysis should be used.

A further disagreement between the informants and the methods concerns the fuzzy entropy of the T-units. The lack of correlation in the levels of fuzzy entropy identified by the informants and the methods challenges the assumption that topics should be treated as fuzzy probabilities. However, the correlation was calculated for the fuzzy entropies of all informants against all methods, and the validity of the topics identified by topical structure analysis and given-new progression is dubious. If we calculate the fuzzy entropies for the methods using only lexical analysis and topic-based analysis, the average fuzzy entropy per T-unit is 0.60 and the correlation with the fuzzy entropies for the informants is 0.22 ($p < 0.05$). Although not high, this significant correlation suggests that fuzzy interpretations should not be discounted.

Perhaps the most noticeable difference between the informants and the methods of analysis concerns the nature of and bases for identifying topic keywords. For the nature of keywords, the most salient difference concerns the inclusion of rhetorical words as parts of the topic by informants whereas the methods of analysis focus almost exclusively on content words. These different types of words reflect two different types of coherence: propositional or ideational coherence and interactional coherence (Lautamatti 1990; Redeker 1990; Sanders et al. 1992). Ideational coherence is based on the semantic or content ties in discourse, whereas interactional coherence is based on illocutionary force or the purposes of discourse which can be described using rhetorical words. Traditionally, *topic* has been considered to be related solely to ideational coherence (and the methods of analysis focus almost exclusively on content), but the identifications of topics by the informants suggest that interactional coherence may also play a role, albeit minor, in topics.

The bases on which topic keywords are identified also differ between the informants and the methods of analysis. While the methods rely solely on the frequency of words, the informants appear to be considering both frequency and saliency (although, without an investigation of how the informants identified topics, the exact basis is unclear). A further distinction concerns the identification of people's names as topic keywords, especially *Gore* (or *I* in the text). The informants identified *Gore* as a keyword rarely, and the T-units in which this was done concerned his role in the story of global warning (e.g. *I wrote a book about it*), in other words, where *Gore* was primarily functioning ideationally. The methods, on the other hand, identified *Gore* as a keyword much more frequently since they considered all instances of *I* in the text, ideational, interactional and textual (e.g. *I would like to emphasize this point*). These different bases are not restricted to analyses of topics – corpus linguistics also tends to focus on frequency at the expense of saliency and to play down functional considerations. It would appear that dealing with saliency and functional considerations are points which quantitative methods in applied linguistics could usefully develop.

To summarise, there are some significant relationships between the findings from the four methods of analysis and the topics identified by the informants, suggesting that the informants' identification of topics and topic boundaries may be influenced to some extent by the textual features analysed by the methods, most notable given-new information and the semantic relationships underpinning topic-based analysis. While significant, the correlations are not noticeably high implying that other issues also influence the informants' identification of topics, an implication supported by the differences in the nature of the keywords identified by the informants and the analyses. The use of paraphrases and shell nouns as keywords by the informants as well as their reliance on saliency as well as frequency suggest issues that text-based analyses of topics need to take into account if they are to reflect how people identify topics.

6. Conclusion

This paper has examined discourse topics as sets of topic keywords and attempted to provide insights into how informants identify topics by comparing the findings from the informants against those of four text linguistic methods of analysing topics. The results show that the most commonly used method of analysing topics, topical structure analysis, is not closely related to how informants identify topics. Rather, if the focus is on topic boundaries, given-new progression is more closely related; and if the focus is on topic keywords, topic-based analysis is preferable. The findings have also highlighted some potential weaknesses in the ways keywords are identified in the methods of analysis. It is hoped that these findings will lead to better-founded analyses of discourse topics in future research.

References

Aktas, Rahime N. and Cortes, Viviana (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes* 7 (1): 3-14.

Berber Sardinha, Antony. P. (1997). Automatic Identification of Segments in Written Texts. Unpublished doctoral dissertation, University of Liverpool.

Brown, Gillian and Yule, George (1983a). *Discourse Analysis*. Cambridge: Cambridge University Press.

Brown, Gillian and Yule, George (1983b). *Teaching the Spoken Language: An Approach Based on the Analysis of Conversational English*. Cambridge: Cambridge University Press.

Carlson, Lauri (1983). *Dialogue Games: An Approach to Discourse Analysis*. Dordrecht: D. Reidel.

Chafe, Wallace L. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In *Subject and Topic*, C. N. Li (ed.), 25-55. New York: Academic Press.

Chafe, Wallace L. (1980). The deployment of consciousness in the reproduction of a narrative. In *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*, Wallace L. Chafe (ed.), 9-50. Norwood, NJ: Ablex.

Clancy, P. M. (1980). Referential choice in English and Japanese narrative discourse. In *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*, Wallace L. Chafe (ed.), 127-202. Norwood, NJ: Ablex.

Connor, Ulla (1996). *Contrastive Rhetoric: Cross-Cultural Aspects of Second-Language Writing*. Cambridge: Cambridge University Press.

Connor, Ulla and Farmer, Mary (1990). The teaching of topical structure analysis as a revision strategy for ESL writers. In *Second Language Writing: Research Insights for the Classroom*, Barbara Kroll (ed.), 126-139. Cambridge: Cambridge University Press.

Coulthard, Malcolm (1977). *An Introduction to Discourse Analysis*. London: Longman.

Covelli, Lucille H. and Murray, Stephen O. (1980). Accomplishing topic change. *Anthropological Linguistics* 22 (9): 382-389.

Crookes, Graham and Rulon, Kathryn A. (1988). Topic and feedback in native speaker/nonnative-speaker conversation. *TESOL Quarterly* 22 (4): 675-681.

Crow, Bryan K. (1983). Topic shifts in couples' conversations. In *Conversational Coherence: Form, Structure and Strategy*, Robert T. Craig and Karen Tracy (eds.), 136-156. Beverley Hills, CA: Sage.

Daneš, Frantisek (1974). Functional sentence perspective and the organization of the text. In *Papers on Functional Sentence Perspective*, Frantisek Daneš (ed.), 106-128. Prague/The Hague: Academia/Mouton.

Davies, F. (1994). From writer roles to elements of text: Interactive, organisational and topical. In *Reflections on Language Learning,* Leila Barbara, Mike Scott and Antonieta Celani (eds.), 170-183. Clevedon: Multilingual Matters.

de Beaugrande, Robert and Dressler, Wolfgang (1981). *Introduction to Text Linguistics*. London: Longman.

Ferret, Olivier and Grau, Brigitte (2000). A topic segmentation of texts based on semantic domains. 14th European Conference on Artificial Intelligence. Berlin August 2000, http://perso.limsi.fr/bg/articles/ECAI2000-OFBG.pdf (accessed 25 September 2008).

Firbas, Jan (1987). On two starting points of communication. In *Language Topics: Essays in Honour of Michael Halliday Volume 1*, Ross Steele and Terry Threadgold (eds.), 23-46. Amsterdam: John Benjamins.

Fries, Peter H. (1983). On the status of theme in English: Arguments from discourse. In *Micro and Macro Connexity of Texts,* Janos S. Petöfi and Emel Sözer (eds.), 116-152. Hamburg: Helmut Buske Verlag.

Fries, Peter H. (1994). On theme, rheme and discourse goals. In *Advances in Written Text Analysis,* Malcolm Coulthard (ed.), 229-249. London: Routledge.

Gardent, Claire and Webber, Bonnie (1998). Describing discourse semantics. Proceedings of the 4th TAG+ Workshop, Philadelphia, http://www.informatik.uni-hamburg.de/WSV/teaching/Unterspez/GardentWebber98 (accessed 25 September 2008).

Gardner, Roderick (1987). The identification and role of topic in spoken interaction. *Semiotica* 65 (1-2): 129-141.

Goldberg, Julia A. (1983). A move toward describing conversational coherence. In *Conversational Coherence: Form, Structure and Strategy*, Robert T. Craig and Karen Tracy (eds.), 25-46. Beverley Hills, CA: Sage.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English. *Journal of Linguistics* 3 (1): 37-81 and 3 (2): 199-244.

Haviland, S. E. and Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. In *Experimenting with the Mind: Readings in Cognitive Psychology*, Lloyd K. Komatsu (ed.), 363-369. Pacific Grove, CA: Brooks/Cole.

Hemphill, L. (1989). Topic development, syntax and social class. *Discourse Processes* 12 (3): 267-286.

Hoey, Michael (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hu, Jianhua and Pan, Haihua (2001). Processing local coherence of discourse in centering theory. Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation. Hong Kong: City University of Hong Kong.

Jäger, Gerhard (2001). Topic-comment structure and the contrast between stage level and individual level predicates. *Journal of Semantics* 18 (2): 83-126.

Keenan, E. Ochs and Schieffelin, B. B. (1976). Topic as a discourse notion: A study of topic in the conversation of children and adults. In *Subject and Topic*, C. N. Li (ed.), 335-384. New York: Academic Press.

Kosko, Bart (1993). *Fuzzy Thinking: The New Science of Fuzzy Logic*. New York: Hyperion.

Langacker, Ronald W. (1996). Conceptual grouping and pronominal anaphora. In *Studies in Anaphora*, Barbara Fox (ed.), 333-378. Amsterdam: John Benjamins.

Lautamatti, Liisa (1978). Observations on the development of the topic in simplified discourse. In *Writing across Languages: Analysis of L2 Text*, Ulla Connor and Robert B. Kaplan (eds.), 87-113. Rowley, MA: Newbury House.

Lautamatti, Liisa (1990). Coherence in spoken and written discourse. In *Coherence in Writing: Research and Pedagogical Perspectives*, Ulla Connor and Ann M. Johns (eds.), 29-39. Alexandria, VA: TESOL.

Lee, Icy (2002). Teaching coherence to ESL students: a classroom inquiry. *Journal of Second Language Writing* 11 (2): 135-159.

Mäkinen, Kaarina (1992). Topical depth and writing quality in student EFL compositions. *Scandinavian Journal of Educational Research* 36 (3): 237-247.

Maynard, Douglas W. (1980). Placement of topic changes in conversation. *Semiotica* 30 (3-4): 263-290.

McCarthy, Michael (1991). *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.

McCarthy, Michael and Carter, Ronald (1994). *Language as Discourse: Perspectives for Language Teaching*. London: Longman.

McNally, Louise (1998). On recent formal analyses of topic. The Tblisi Symposium on Logic, Language and Computation, http://mutis.upf.es/~mcnally/tbilisi.pdf (accessed 25 September 2008).

Palmer, F. R. (1981). *Semantics*, second edition. Cambridge: Cambridge University Press.

Redeker, Gisela (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14 (3): 367-381.

Richards, Jack C. and Schmidt, Richard W. (1983). Conversational analysis. In *Language and Communication*, Jack C. Richards and Richard W. Schmidt (eds.), 117-154. London: Longman.

Rost, Michael (1994). *Introducing Listening*. London: Penguin.

Rutherford, William E. (1987). *Second Language Grammar: Learning and Teaching*. London: Longman.

Sanders, T. J. M., Spooren, W. P. M. and Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes* 15 (1): 1-35.

Schneider, Melanie and Connor, Ulla (1990). Analyzing topical structure in ESL essays: Not all topics are equal. *Studies in Second Language Acquisition* 12 (4): 411-427.

Scott, Mike (1997). PC analysis of key words – and key key words. *System* 25 (2): 233-245.

Shepherd, John D. (1998). Storytelling in Conversational Discourse: A Collaborative Model. Unpublished doctoral dissertation, University of Birmingham.

Stech, E. L. (1982). The analysis of conversational topic sequence structures. *Semiotica* 39 (1-2): 75-91.

Tomlin, Russell S., Forrest, Linda, Pu, Ming Ming and Kim, Myung Hee (1997). Discourse semantics. In *Discourse as Structure and Process:*

Discourse Studies: A Multidisciplinary Introduction Vol. 1, Teun A. van Dijk (ed.), 64-111. London: Sage.

van Dijk, Teun A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. London: Longman.

Walker, Marilyn A., Joshi, Avarind K. and Prince, Ellen F. (1998). Centering in naturally-occurring discourse: An overview. In *Centering in Discourse*, Marilyn A. Walker, Avarind K. Joshi and Ellen F. Prince (eds.), 1-28. Oxford: Clarendon Press.

Watson Todd, Richard (1998). Topic-based analysis of classroom discourse. *System* 26 (3): 303-318.

Watson Todd, Richard (2003) Topics in classroom discourse. Unpublished doctoral dissertation, University of Liverpool.

Watson Todd, Richard (2005) A fuzzy approach to discourse topics. *Semiotica* 155 (1-4): 93-124.

Wikborg, Eleanor (1990). Types of coherence breaks in Swedish student writing: Misleading paragraph division. In *Coherence in Writing: Research and Pedagogical Perspectives*, Ulla Connor and Ann M. Johns (eds.), 131-148. Alexandria, VA: TESOL.

Witte, Stephen P. (1983). Topical structure and revision: an exploratory study. *College Composition and Communication* 34 (3): 313-341.