A Collocation Analysis of Topic Change Utterances in Multi-party Meeting Conversations

Shunichi Ishihara
The Australian National University
Shunichi.Ishihara@anu.edu.au

Abstract. The aim of this paper is to quantitatively investigate the sorts of expressions that are used around the boundaries of conversation topics in multi-party meeting conversations. This will be done by means of identifying word collocations which are statistically-significantly associated with topic boundaries, and graphically presenting them in the form of a network. The International Computer Science Institute Meeting Recorder Dialogue Act Corpus is used in this study. We will demonstrate that the derived network of collocated words supports previous studies undertaken in the area of conversation analysis on topic changes in some respects, while other findings of conversation analysis are not confirmed by the empirical results of our study. This study presents expressions which seem to be distinctive to multi-party meeting conversations by referring to the nature of meeting conversations

Keywords: topic change, collocation, multi-party meeting, conversations

Selected Papers from the 2009 Conference of the Australian Linguistic Society,
edited by Yvonne Treis & Rik De Busser. http://www.als.asn.au. 2010.

1. Introduction

Previous work on topic changes in conversation analysis and automatic topic segmentation has found that cue phrases, such as *okay*, *anyway*, and *alright*, provide valuable information regarding the structure of discourse (Grosz & Sidner 1986, Howe 1991, Passonneau & Litman 1997). Cue phrases are "words and phrases that directly signal the structure of a discourse" (Hirschberg & Litman 1993:510). Cue phrases here may be termed differently (with or without some differences in definitions and groupings) depending on the opinion of the scholars, these include the terms "acknowledgement tokens" (Howe 1991), "prefatory discontinuity markers" (Drew & Holt 1998), "discourse particles" (Galley & McKeown 2003), etc.

In the area of conversation analysis, techniques for achieving topic changes have been described in detail (Maynard 1980, Orletti 1989, Howe 1991, Geluykens 1993, Drew & Holt 1998). These techniques will be summarised in the following subsection. Although these findings are very useful, they unfortunately lack a statistical and mathematical foundation as conversation analysis concentrates on the local management of specific aspects of conversation. Therefore, even if it is reported, for example, that the formulaic expression *speaking of* is often used to introduce a new topic, we do not know whether this formulaic expression is statistically-significantly associated with topic changes. That is, how significantly (or how strongly) this expression is correlated with topic changes.

In some topic segmentation algorithms, utterance-initial cue phrases (and sometimes other lexical cues) are used together with other topic-change indicators to detect topic boundaries (Galley & McKeown 2003). However, the findings of conversation analysis are not extensively reflected in topic segmentation algorithms. This may be partly due to the above-mentioned lack of a statistical foundation.

This study attempts to fill in the gap between these two areas of study. More precisely, we will identify word collocations – which are statistically-significantly associated with topic changes – from the utterances around topic boundaries, and represent them in the form of a network. In this study, the term *network* refers to inter-relations or associations of words (Aitchison 1994:82). These collocations and their network enable us to quantitatively investigate the types of expressions that are

involved in the utterances around topic boundaries. The observations made throughout this study are useful to the linguistic understanding of the mechanisms of topic changes by providing statistical proof, while the same results are also easily applicable to automatic topic segmentation.

In this study, both grammatical and lexical single words that have statistical significance in relation to topic boundaries were extracted in order to investigate their co-occurrence patterns. These single words are referred to as "cue words" in this study.

Almost all previous studies on topic changes in conversation analysis are based on two-party natural conversations (e.g. a conversation between two friends). However, with the application of automatic topic segmentation in mind, we decided to investigate multi-party meeting room conversations in this study. The findings reported from previous studies on topic changes in two-party conversations will be compared with the findings of the current study.

1.1. Techniques for achieving topic changes

Daily conversations commonly consist of various transitions from one topic to another. However, topic changes are not random events but are well controlled in terms of when and where they take place (Maynard 1980:264). This subsection summarises previous studies reporting how topic changes (including both the ending and the beginning of a topic) are achieved. We focus on the following six commonly perceived techniques:

- Prefatory disjunctives
- Questions
- Declarative clauses
- Explicit topic change expressions
- Formulaic expressions
- Summary assessment

Prefatory disjunctives, which are very similar to what Hirschberg & Litman (1993:501) call "cue phrases", are discontinuity markers such as *anyway*, *alright*,

well, *okay*, etc. They are disjunctives in that "they work to disengage the forthcoming turn from being tied or connected to, or coherent with, its prior turn" (Drew & Holt 1998:510). A large number of studies impressionistically or quantitatively report that these small phrases are associated with topic changes (Grosz & Sidner 1986, Hirschberg & Litman 1993, Passonneau & Litman 1997).

It has been reported that questions, whether yes-no questions or wh-questions, are often used to introduce a new topic (Howe 1991, Geluykens 1993).

Simple declarative clauses are also used to initiate new topics. Geluykens (1993:204) states that this is the least obtrusive way of introducing a new topic. Geluykens (1993:197-199) specifically reports cases in which a new topic is introduced by what he calls "existential *there* clauses", such as *There is this guy I know, he likes Mary*.

Although this technique is said to be rarely used in conversation, there are some situations in which speakers directly introduce a new topic by using explicit expressions, such as *I'll tell you something else…*, *now switching to…* or *another topic* (Geluykens 1993:209).

According to Howe (1991:94-96), formulaic expressions, such as *talking about*, *speaking of*, *this reminds me* and *excuse me*, are used as topic-beginning indicators.

Summary assessment is commonly used to conclude a topic (Howe 1991:77-78). Summary assessment is a comment, usually taking the form of a statement, on the preceding topic which seems to close off the topic from further discussion. It may also function as a formulation, clarifying the central point of the topic or stating the consequence of what has been talked about. Therefore, summary assessment adds little, if any, new information to the preceding topic (Howe 1991:77). Although Howe does not explicitly mention this, many of those utterances which she lists as examples of summary assessment contain words of value judgement, such as *interesting*, *wonderful*, *dreadful*, *good*, *great*, *well* and so on.

Drew & Holt (1998) report that figurative expressions often appear as the summary of a topic to manage topic transition in conversations.

As well as the verbal techniques introduced above, non-verbal cues are also employed to mark a new topic. These non-verbal cues include: an audible intake of

breath, laughter, pause, raised pitch, etc (Howe 1991, Drew & Holt 1998). These techniques, however, are beyond the scope of the current study.

2. Database

The ICSI (International Computer Science Institute) Meeting Recorder Dialogue Act (MRDA) Corpus (Shriberg et al. 2004) has been used in this study. The MRDA is a hand-annotated version of the ICSI Meeting Corpus (Morgan et al. 2001) which contains 75 naturally occurring multi-party meetings, each approximately one hour in length. 53 different speakers appear in the corpus, with an average of approximately six speakers per meeting. A stream of dialogue is segmented in terms of utterances, each of which constitutes prosodically one unit. The annotation provides three types of information: marking of dialogue act (DA), marking of DA segment boundaries, and marking of correspondences between DAs (= adjacency pairs). An example of the MRDA Corpus is given in Table 1 (SP = speakers, DA = dialogue acts, AP = adjacency pairs). The comprehensive explanations of the MRDA Corpus, including various kinds of tags, can be found in Dhillon et al. (2004).

Time	SP	DA	AP	Transcript
442.938-447.028	c3	s	25b.26a	it's ics- - uh icsi has a format for frame-level representation of features .
447.808-448.338	cB	s^bk	26b	o_k .
448.22-448.67	c3	fh		um ==
448.388-452.688	cB	s^bu	26b+.27a	that you could call - that you would tie into this representation with like an i_d .
451.177-451.527	c3	s^aa	27b	right .
452.755-453.065		s^aa^r	27b+	right .
453.255-457.595	c3	s	27b++.28a.29a	or - or there's a - there's a particular way in x_m_l to refer to external resources .
453.742-454.122	cB	fh		and ==
457.809-458.249	cB	s^bk	28b	o_k .
458.453-461.423		s:s^co	27b+++.29a+	so you would say refer to this external file .

Table 1. An example of the MRDA Corpus

Shriberg et al. (2004) examined the reliability of three labellers for both segmentation and DA labelling using κ-statistics, and confirmed that the agreement between the labellers is appropriate for this type of task ($\kappa = 0.80$).

Scholars have varying definitions for conversation topics (Maynard 1980, Geluykens 1993, Orletti 1989). The following is the labelling guideline for topic change locations (tc) (Dhillon et al. 2004:100) used in the MRDA corpus.

> The <tc> tag marks utterances which either begin or end a topic. As the <tc> tag marks when a topic changes, once the topic has indeed changed and a new topic is in the course of discussion, the discussion of the new topic is not marked with the <tc> tag.
>
> Oftentimes, a speaker will utter a floor grabber <fg> and then introduce a new topic. As the floor grabber appears as though it is used as a mechanism to gain the floor and introduce a new topic, and in effect signals a change in topic, it is not marked with the <tc> tag. Rather, only utterances which convey a change in topic are marked with the <tc> tag. In which case, a speaker must specify in his utterance that he wishes to end a topic or else he must state that he wishes to begin a new topic either by initiating and specifying a new topic or else by merely stating that he wishes to talk about something else.

679 topic change locations (tc) are annotated in the MRDA Corpus.

3. Word collocation: Methodology

In this study, utterances which occur within approximately ten seconds before and after a topic change DA (tc) are identified as topic change utterances. It is "approximately" ten seconds as speech is continuous, and utterances do not always finish or start at precisely ten seconds before or after a topic change location. Instances in which an utterance was uttered at ten seconds before or after a topic change location have also been included in the analysis. These topic change utterances are those which are analysed for word collocations. However, it is necessary to refer to all utterances which appear in the entire database in order to identify word collocations which are significantly associated with topic changes.

The following five steps are taken in this study to identify word collocations that are statistically-significantly associated with topic changes.[1]

Step 1: Remove punctuation/diacritic marks and unwanted words (e.g. incomplete words) from the transcribed texts, and tokenise the resultant clean texts.

Step 2: Make a list of different words, and remove low frequency words from the list.

Step 3: Identify cue words that are significantly associated with topic boundaries.

Step 4: Use the cue words identified in Step 3 to find collocations that are significantly associated with topic boundaries.

Step 5: Remove pseudo-collocations from the collocations identified in Step 4 to draw a network of true collocations.

These five steps are elaborated in the following subsections.

3.1. Step 1

As can be seen in Table 1, transcribed texts may contain diacritic and punctuation marks, such as "= =", "-" and incomplete words and so on. These were removed from the transcribed texts before tokenisation. Stemming algorithm was not applied in Step 1. That is, for example, *talked* and *talk* are treated as different words.

3.2. Step 2

The frequencies of the tokenised words from Step 1 were calculated. Following Step 1, the total number of words was 732,918, and the number of different words was 12,591. These were from the entire database. Low frequency words were removed for further analysis because 1) these low frequency words may have appeared in the database by chance due to the specific nature of topics and 2) we would like to focus

[1] Readers with little background in mathematics and statistics are advised to read chapter five of Manning & Schütze (1999), in which they explain the statistics that are available and how they can be used for the analysis of word collocations.

on those words which are generally used regardless of the topic. Although 10, 50 and 100 were arbitrarily set as thresholds in the original study; results based on a threshold of 50 are given in this paper. By setting 50 as the threshold, 1,035 different words were selected as a result of Step 2.

3.3. Step 3

Steps 3 and 4 are designed for word collocations around topic boundaries. Step 3 involves the identification of those cue words which show a correlation with topic boundaries. As explained above, utterances which occur within approximately ten seconds before and after a topic change DA (tc) are chosen as topic change utterances. To identify these cue words, $Yates'\chi^2$ test was employed in this study. $Yates'\chi^2$ test is essentially the same as χ^2 test, but it is adjusted in such a way as to prevent overestimation of the statistical significance for small sampled data. $Yates'\chi^2$ value is calculated using the formula defined in (1).

$$Yates'\,\chi^2 = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (1)$$

$Yates'\chi^2$ test is based on a 2x2 contingency table which shows the frequencies of occurrence of all combinations of the levels of two dichotomous variables, in a sample of size N (= 1035 at Step 2). In Step 3, the two dichotomous variables are a given word w_i $(i = 0, 1 \ldots N-1)$ and *topic boundary* (*TB*). The combinations of the levels of these two variables are (w_i, TB), $(\sim w_i, TB)$, $(w_i, \sim TB)$ and $(\sim w_i, \sim TB)$ as can be seen in Table 2.

	w_i	$\sim w_i$	total
TB	a	b	a+b
~TB	c	d	c+d
total	a+c	b+d	n (= a+b+c+d)

Table 2. 2x2 contingency table for $Yates'\chi^2$ test (Step 3)

In Table 2, a, b, c and d are the frequencies of the occurrence of (w_i, TB), $(\sim w_i, TB)$, $(w_i, \sim TB)$ and $(\sim w_i, \sim TB)$, respectively. Using the word w_i as an example, a is the frequency of w_i appearing in topic change utterances; b is the frequency of all words, except for w_i, appearing in topic change utterances; c is the

frequency of w_i appearing in non-topic change utterances; and d is the frequency of all words, except for w_i, appearing in non-topic change utterances.

Using Table 2 and Formula (1), those words for which the $Yates'\chi^2$ value rejected the hypothesis under a 0.005-level of confidence (the rejection criterion is $\chi^2 \geq$ 7.8794) were selected as cue words ($W^{TB} = \{W_0^{TB}, W_1^{TB} \ldots W_{N-1}^{TB}\}$, $N =$ the total number of cue words in Step 3). For the word w_i, for example, if we obtained 10, 300, 2 and 500 for a, b, c and d, respectively, $Yates'\chi^2$ value is 8.6703 ($= \frac{812*(|10*500-300*2|-{}^{812}/_{2})^2}{(10+300)*(2+500)*(10+2)*(300+500)}$). The $Yates'\chi^2$ value of 8.6703 indicates that the word w_i is significantly associated with topic boundaries.

The total number of cue words obtained in Step 3 was 139. Table 5 contains the 15 cue words with the highest $Yates'\chi^2$ values.

3.4. Step 4

Step 4 investigates the dependency of any two cue words in W^{TB} in topic change utterances. That is, all possible combinations of two cue words (W_i^{TB}, W_j^{TB}) of W^{TB} (refer to the matrix given in Table 3) were tested in terms of their dependency ($Yates'\chi^2_{i,j}$) using the contingency table given in Table 4. The number of all possible combinations of two cue words was 9,870.

	$\boldsymbol{w_0^{TB}}$	$\boldsymbol{w_1^{TB}}$	$\cdots$	$\boldsymbol{w_{N-1}^{TB}}$
$\boldsymbol{w_0^{TB}}$	$Yates'\chi^2_{(0,0)}$			
$\boldsymbol{w_1^{TB}}$	$Yates'\chi^2_{(0,1)}$	$Yates'\chi^2_{(1,1)}$		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
$\boldsymbol{w_{N-1}^{TB}}$	$Yates'\chi^2_{(0,N-1)}$	$Yates'\chi^2_{(1,N-1)}$	$\cdots$	$Yates'\chi^2_{(N-1,N-1)}$

Table 3. A matrix showing all possible co-occurrence patterns of cue words (N = 139)

In Table 4, a, b, c and d are the frequencies of the occurrence of (w_i^{TB}, w_j^{TB}), $(\sim w_i^{TB}, w_j^{TB})$, $(w_i^{TB}, \sim w_j^{TB})$ and $(\sim w_i^{TB}, \sim w_j^{TB})$, respectively ($i = j = \{0, 1, 2 \ldots 138\}$) in topic change utterances. Supposing that there are two words w_i and w_j in the selected 139 cue words, a is the frequency with which the words w_i and w_j occur together in the same topic change utterances; b is the frequency with which w_i and non-w_j words occur together in the same topic change utterances; c is the frequency with which non-w_i words and w_j occur together in the same topic

change utterances; d is the frequency with which non-w_i and non-w_j words occur together in the same topic change utterances. Using Table 4 and Formula (1), the two cue words with $Yates'\chi^2$ value which rejected the hypothesis under a 0.005-level of confidence were selected as the collocations significantly associated with topic change. 660 collocations were derived in Step 4.

	w_i^{TB}	$\sim w_i^{TB}$	total
w_j^{TB}	a	b	a+b
$\sim w_j^{TB}$	c	d	c+d
total	a+c	b+d	n (= a+b+c+d)

Table 4. 2x2 contingency table for $Yates'\chi^2$ test (Step 4)

3.5. Step 5

Manning & Schütze (1999: 151) define collocation as "an expression consisting of two or more words that correspond to some conventional ways of saying things". 660 collocations were identified in Step 4 as being significantly associated with topic boundaries. However, if one tries to graphically present a network of the 660 collocations derived from Step 4, the network becomes too busy visually to comprehend due to the large number of associations between words. Furthermore, as we obtained these collocations on the basis of utterances of varying lengths, there is a possibility that the 660 collocations may include some pseudo-collocations which are not consistent semantically and/or morpho-syntactically in terms of the co-occurrence pattern of two cue words. As the aim of this paper is to investigate the sorts of expressions that are involved in topic change utterances from word collocations and their network, we need to eliminate these pseudo-collocations.

We eliminated pseudo-collocations by calculating the mean (μ) and the standard deviation (sd) of the distance between collocated cue words ($|W_i^{TB} - W_j^{TB}|$) in topic change utterances. That is, if the μ of any given collocated cue words is large – meaning they do not appear closely to each other – those two cue words are less likely to be semantically and/or morpho-syntactically cohesive. Likewise, if the sd of any given collocated cue words is large, there is no consistent co-occurrence pattern, and they do not form a fixed pattern.

By observing the point when the network starts making sense conceptually by changing the cut-off parameters, a μ and an sd of 4.5 were set as the cut-off

parameters to eliminate pseudo-collocations in this study. These cut-off parameters appear to be appropriate in light of the distributions of the μ and the sd values of the 660 collocations. The μ and sd of the distance $(|W_i^{TB} - W_j^{TB}|)$ of each of the 660 collocations in topic change utterances are plotted in Figure 1a and Figure 1b, respectively, in ascending order, along the x-axis.

In Figure 1a, the number of collocated words decreases at more or less a constant rate between the μ values of 10 and 5. This rate starts changing between the μ values of 5 and 4 (refer to the arrow in Figure 1a). In other words, the steepness of the slope changes around the μ values of 5 and 4. Similarly, in Figure 1b, after a constant decrease rate between the sd of 10 and 5, this rate starts changing between the sd of 5 and 4 (refer to the arrow of Figure 1b). These observations indicate that there is a difference in nature amongst the 660 collocations, with a μ and a sd of 4.5 serving as a boundary.

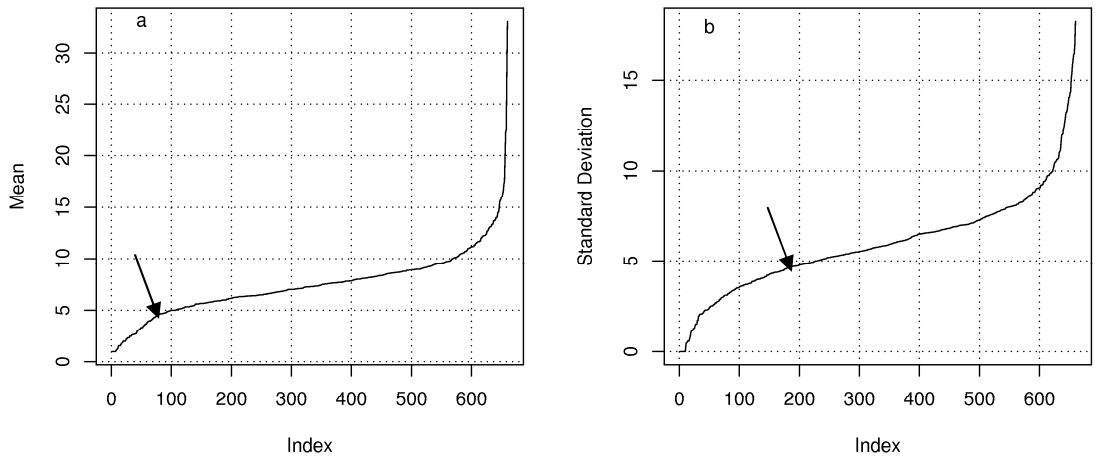

Figure 1. The mean (a) and the standard deviation (b) of the distance of two collocated words (660) plotted in ascending order

By setting the cut-off parameters accordingly, the 68 collocations given in Table 6 were selected as genuine collocations. A graph drawing package developed by AT&T called NEATO (Ellson et al. 2005) was used to graphically present the inter-relationship of these collocations in the form of a network. The network is given in Figure 2.

4. Results and discussions

In this section, the results for the cue words are discussed first, and then those for the collocations of the cue words are discussed by referring mainly to the resultant network.

The total number of cue words obtained in Step 3 was 139. Table 5 contains the 15 cue words with the highest $yates'\chi^2$ values. Many of the words given in Table 5 are general words which are used regardless of the types of topics. Furthermore, many of them match those reported in previous works (i.e. *o_k*, *anyway*, *alright*, *let's*, etc). Particularly, *o_k* shows the highest $yates'\chi^2$ value ($= 499.4$) of all, a value which is significantly higher than the rest. In many of the previous works on automatic topic segmentation, only cue phrases appearing at the beginning of utterances or sentences were considered (Passonneau & Litman 1997, Galley & McKeown 2003). However, in this study, the selection of cue words is not limited to words in the utterance- or sentence-initial positions. Consequently, grammatical words, such as *been* and *about*, and adverbs, such as *else*, were selected as significant cue words. As will be shown below, these words occur with other cue words, and they are significantly correlated with topic boundaries.

w^{TB}	$Yates'\chi^2$
o_k	499.4
agenda	154.2
talk	134.2
about	132.7
alright	129.6
anyway	110.6
let's	105.7
so	103.5
uh	92.3
been	74.3
um	73.1
yeah	67.7
you	59.9
last	56.5
else	54.1

Table 5. Cue words

Table 6 (μ and sd = mean and standard deviation of the distance between w_i^{TB} and w_j^{TB}; C = count) contains the 68 collocations selected in Step 5. It is evident from Table 6 that many of the 68 collocations are semantically and/or syntactically self-explanatory, creating formulaic expressions, e.g. the highest 5 collocations (*that'd*, *great*), (*anything*, *else*), (*go*, *ahead*), (*i've*, *been*), (*let*, *me*).

	w_i^{TB}	w_j^{TB}	χ^2	μ	C	sd		w_i^{TB}	w_j^{TB}	χ^2	μ	C	sd
1	that'd	great	810	2.0	6	0.0	35	let's	o k	13	1.9	12	1.1
2	anything	else	671	1.0	25	0.0	36	guess	i	13	1.6	82	2.1
3	go	ahead	582	1.0	17	0.0	37	yeah	um	13	2.6	21	2.2
4	i've	been	406	1.5	34	1.5	38	let	let	12	3.0	3	1.7
5	me	let	188	1.1	17	0.4	39	you	talk	12	4.1	15	2.9
6	oh	yeah	141	1.8	32	1.9	40	why	i	12	3.1	13	2.3
7	about	talk	131	2.6	72	3.7	41	had	i	12	3.4	48	3.6
8	thank	you	111	1.0	12	0.0	42	but	it's	12	3.1	30	3.3
9	playing	i've	67	2.0	5	0.0	43	working	been	12	1.5	7	1.1
10	next	week	52	1.0	10	0.0	44	yeah	so	12	2.3	36	2.7
11	oh	o_k	47	1.3	19	0.8	45	it	been	12	3.6	5	2.0
12	week	last	47	2.0	11	3.6	46	else	uh	11	2.5	4	1.7
13	playing	been	39	1.0	5	0.0	47	next	you	11	4.2	4	2.2
14	on	working	38	2.3	23	2.6	48	it's	so	11	3.3	28	3.9
15	ahead	i'll	34	3.3	3	2.3	49	that	been	11	3.9	18	2.6
16	that'd	good	33	2.3	3	0.5	50	is	meeting	11	4.0	10	3.2
17	mean	i	31	2.2	129	3.7	51	oh	so	10	3.1	7	2.3
18	should	we	26	2.6	72	4.2	52	um	o_k	10	2.7	19	1.9
19	better	it's	24	4.2	5	4.0	53	a	talk	10	3.5	16	2.5
20	mention	wanted	21	2.5	4	1.0	54	working	i've	10	2.8	5	1.3
21	done	we're	20	1.5	11	0.6	55	list	agenda	10	3.3	3	4.0
22	thanks	o_k	20	2.6	3	1.5	56	move	should	10	2.0	4	1.4
23	of	here	20	4.4	15	3.5	57	meeting	know	9	4.3	3	4.0
24	it's	about	19	3.7	4	3.2	58	thing	other	9	1.7	33	3.3
25	send	i'll	18	1.0	4	0.0	59	was	yeah	9	4.0	12	3.6
26	it's	great	18	1.1	6	0.4	60	ahead	o_k	9	4.0	4	4.0
27	is	mean	17	4.3	11	4.1	61	did	we	9	2.7	18	2.1
28	but	but	16	2.5	4	0.5	62	on	talk	9	4.4	5	3.2
29	talked	about	15	2.0	18	2.4	63	since	that	9	2.6	3	2.0
30	but	anyway	15	1.5	14	1.2	64	do	let's	9	2.6	20	3.4
31	eh	uh	14	1.2	4	0.5	65	do	should	9	3.9	34	3.8
32	a	couple	14	2.2	33	2.9	66	let's	alright	8	2.3	3	1.5
33	move	on	14	1.0	8	0.0	67	what's	the	8	3.2	13	2.3
34	let's	let's	14	3.0	6	1.2	68	wanted	you	8	3.3	9	3.0

Table 6. 68 collocations selected in Step 5

The inter-relationships of the 68 collocations given in Table 6 are graphically presented in the form of a network in Figure 2. Note that there are some differences in the length of connecting lines and also that some edges overlap (i.e. *wanted* and *thanks*) in Figure 2. These are due to the technical and graphical issues that emerge as the software tries to draw a graph in the optimal way.

Figure 2 shows one large cluster and eight small clusters comprised of 68 nodes (collocations) and 69 edges (arcs connecting two nodes). There are some pivotal nodes (cue words) which have more than one edge, such as *i*, *o_k*, *it's*, *been*, *talk*, etc. The network given in Figure 2 allows us to identify various expressions that consist of multiple cue words, which are statistically-significantly associated with topic changes.

By analysing naturally-occurring two-party conversations using the methodology of conversation analysis, Howe (1991) reports that a so-called summary assessment often appears at the end of a topic. Summary assessment is characterised as an utterance contributing little, if any, new information to the topic concerned. Although there is a high possibility that some of the collocations make up part of a summary assessment, from the network given in Figure 2, it is difficult to see if there are any particular patterns which can be significantly associated with summary assessment. However, some expressions of value judgement or assessment, such as (*that'd*, *great/good*), (*it's*, *great*), etc. – which are likely to be used by the chair of a meeting – can be identified from the collocations. This point conforms to the examples of summary assessment given by Howe (1991). The collocation (*that'd*, *great*) has the highest $Yates'\chi^2$ value of 810.

In the current data, the collocations (*what's*, *the*) and (*anything*, *else*) are mostly used as questions. Geluykens (1993) and Howe (1991) note that questions (both yes-no and wh-questions) often initiate new topics. However, our results show that only 'what' questions (e.g. *what's the difference between t_l_d_a and s_l_d_a?; so what's the other thing on the agenda actually?*) (not 'how', 'where', etc.) are significantly associated with topic changes. The collocation (*anything*, *else*) is the second strongest topic boundary indicator ($yates'\chi^2 = 671$), and is constantly used as a question mainly by the chair. The collocations, such as (*talk/talked*, *about*, *a*, *couple*),

(*talk*, *on*), (*wanted*, *mention/talk*), often appear as question sentences uttered by a chair in the current database (e.g. *you wanna talk about recognition?*).[2]

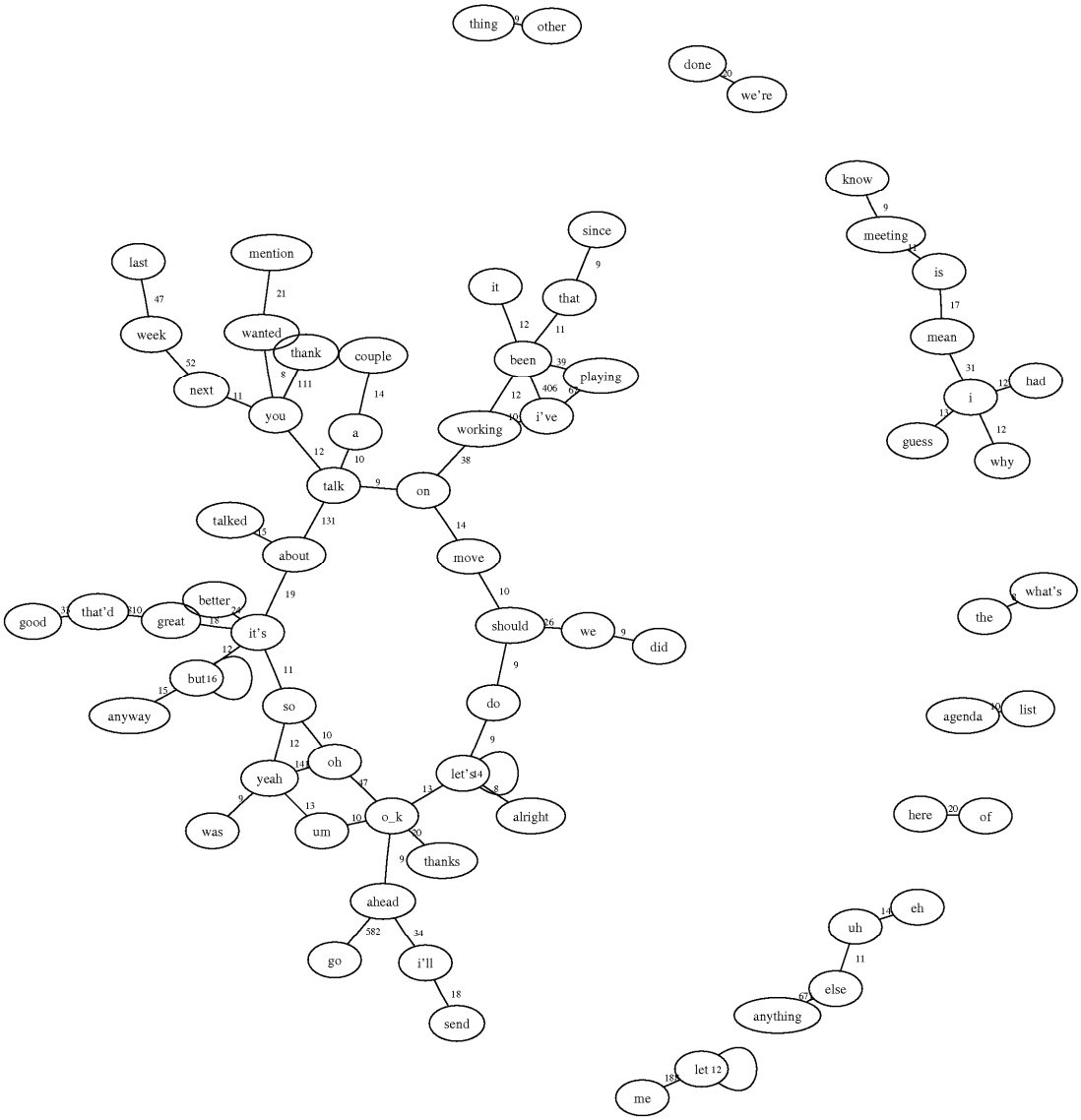

Figure 2. Network of identified 68 collocations

[2] Note that following the original transcribed texts of the MRDA corpus, example utterances are given all in lower case in this study.

As Geluykens (1993:204) reports, some collocations indicate that declarative clauses are used to introduce a new topic. Some collocations, such as (*i've*, *been*, *working/playing*, (*on*)), are used to report on progress made, or an action taken on a topic (e.g. *so i've been uh working still on the spectral subtraction*; *i've been exploring a parallel v_a_d without neural network*). The collocation (*we*, *did*) is also frequently used to report what was done on a topic (e.g. *this is the things that we did in the last three months*). The collocation (*last*, *week*) is often accompanied with the utterances of reporting (e.g. *i spent the last week understanding some of the data*; *so what happened since um last week is…*). A similar collocation is (*next*, *week*), which is used to talk about the action that will be taken on a topic (e.g. *and then continue with this next week*). In the same line, the collocation (*i'll*, *send*) is also used as an action which will be taken on a topic (e.g. *so tomorrow i'll send a email and just ask if he received it*; *i'll send mail to speech local and see if anyone's still using it*).

Unlike the note made by Geluykens (1993:197-199) about "existential *there* clauses", the collocation given in Table 6 does not contain *there* as a member of a collocation.

As Geluykens (1993:208-210) reports, in our data, there are some collocations, such as (*o_k/alright*, *let's*, *do*), (*we*, *should*, *do/move*) and (*let*, *me*), which are used to influence the course of conversations (e.g. *well let's do the first one*; *should we move on the technical side*; *so let me suggest we switch to another one*). However, unlike Geluykens' (1993: 209) remark on explicit topic change expressions in which he states that these expressions are rarely used, these topic-change expressions are frequently used in the current database (as shown in Table 6). The rareness of these explicit expressions in two-party conversations may be due to "the collaborative nature of topic change" (Howe 1991:125-128). Geluykens (1993:210) acknowledges that explicit topic change expressions probably appear more often in other discourse types. Perhaps the frequent use of explicit topic change expressions in the current data is another difference between two-party conversations and multi-party meeting conversations. In order to carry out multi-party meetings efficiently, these expressions often need to be used (but not always) by the chair.

The verbs *talk* and *mention*, which are semantically similar, have complicated collocational relationships with other cue words, as can be seen in the collocations (*talk/talked*, *about*, *a*, *couple*), (*talk*, *on*), (*you*, *wanted*, *mention/talk*), etc. Depending

on the subject of an utterance, the actual utterances in which these collocations occur have two main functions. Partly relating to explicit topic change expressions, if the subject of the sentence is the speaker (first person singular), then the speaker may use them to get the conversation floor and *talk about* something (e.g. *so other topics i wanted to talk about are…*). This kind of overt attempt of obtaining a conversational floor appears to be rather unique to multi-party meeting conversations. If it is uttered by the chair of a meeting and the subject of the utterance is a second person, the chair may use them to induce a given participant to *talk about* a given item (e.g. *you had, you wanted to talk about the…*).

The first person singular pronoun has four edges making up the collocations of (*i*, *guess*), (*i*, *mean*), (*i*, *had*) and (*i*, *why*). These collocations are used in statements. The collocation (*i*, *guess*) is used in statements providing some sort of opinion, desire, etc. (e.g. *there's i guess a little more wiring to do…*; *i guess i would like to have a discussion about…*; *i guess it depends really*). The collocation (*i*, *mean*) is also used in statements (e.g. *i mean i don't say anything about where you live on the form*; *i mean it's different*). It is not immediately obvious why this is so, but it is interesting that these functions are related to topic changes. The collocations (*i*, *guess*) and (*i*, *mean*) are very frequently used in the current database (counts: 82 and 129, respectively). The collocation (*i*, *had*) is used in various kinds of statements, but it appears that the collocation is often used in the sentences which provide a general introductory statement before getting into the core of a new topic (e.g. *i had a question for adam*; *i had something that i could bring up*; *that reminds me i had a thought of an interesting project*). The collocation (*i*, *why*) is used in statements where a speaker would like to provide a reason for his/her action (e.g. *that's why i said point to robert when i did it*; *i mean that's why i thought about it*).

As briefly commented above, one of the characteristics of multi-party meeting conversations is the involvement of a chair whose remarks can strongly influence the direction/progress of a meeting. The collocation ((*o_k*), *go*, *ahead*), which has a very high $Yates'\chi^2$ value ((*go*, *ahead*) = 582), is, for the most part, limited to the chair of a meeting in the current data, and unique to multi-party meeting conversations.

It is also clear from the collocations that expressions showing appreciation, such as (*thank*, *you*) and (*o_k*, *thanks*), are significantly associated with topic changes. These expressions tend to be uttered by the chair of a meeting in the current database.

The collocation (*it's*, *better*) which is used in the expression of a suggestion is mainly used in this database to give a suggestion to a given problem (e.g. *maybe it's better to wait*; *it's better to edit out every time you bash microsoft*).

The collocation (*agenda*, *list*) appears in Figure 2. This collocation is fairly self-explanatory in its significant association with topic changes in meeting conversations. This is because meetings usually have an agenda and they progress by referring to it.

The collocation (*we're*, *done*) is used in the expressions by which speakers confirm the end of a topic or even a meeting (e.g. *so we're done with the topic*; *well i guess we're about done*). This sort of utterance, which expresses relief that the meeting has been progressing or has finally ended, may be common with meeting situations, but is considered to be rare in two-party daily conversations.

The small particles, such as *so*, *o_k*, *eh*, *uh*, *um*, *yeah*, etc. – which are used in many cases as a discourse marker – very frequently co-occur with each other. The cue words of *let's*, *let* and *but* are repeated once or twice (or even more) in order to maintain or get one's conversation floor (e.g. *so uh let's let's do our let's do our digits*; *so let me let me just uh finish you know*; *but but uh i i'd never heard that before*).

5. Conclusion and further studies

In this paper, we have demonstrated by means of word collocations that there are many types of expressions involved in topic changes. These expressions have various functions, such as judgement, assessment, reporting, influence to the course of a conversation, management of conversation floor, confirmation of the end of a topic, and so on. We have also argued that some expressions, such as the explicit expressions of influencing the course of conversations and overt expressions with which a speaker attempts to gain a conversational floor, are unique to multi-party meetings, as they have not been reported in previous studies based on naturally

occurring two-party conversations. One of the important observations is that the explicit topic change expressions are frequently used in multi-party meeting conversations, such as (*o_k/alright*, *let's*, *do*), (*we*, *should*, *do/move*) and (*let*, *me*). The expressions uttered by the chair of a meeting – which are also particular to meeting conversations – tend to have high $Yates'\chi^2$ values (e.g. (*that'd*, *great*), (*anything*, else), (*go*, *ahead*), (*thank*, *you*)). The identification of the differences between two- and multi-party conversations is important from a discourse-typological point of view (Geluykens 1993:210).

Based on our findings, we plan to investigate if there are any expressions which are specifically correlated with the opening or ending of a topic in the future. It will also be interesting to see how collocational information, as opposed to cue words, improves the performance of automatic topic segmentation systems. This point is also related to the cut-off parameters, such as low frequency words, the μ and sd of $|w_i^{TB} - w_j^{TB}|$, which were arbitrarily set in this study. These parameters can be empirically set judging from the performance of automatic segmentation systems.

Sacks (1992:15-16) remarks on a stepwise transition between topics. Along the same line as Sacks, Howe (1991:122-124) says that a topic change is a collaborative activity of participants over a certain period of time, in which several topic change indicators seem to appear. These statements seem to be intuitively true by observing sequences of utterances from the current database. However, as we investigated word collocations synchronically within an utterance, we do not know quantitatively what sort of topic change collocations tend to co-occur diachronically. This point is important as it allows one to understand the true mechanism of topic changes.

Acknowledgements

This study was financially supported by the College of Asia and the Pacific, the Australian National University. The author is very grateful to Dr Tim Hassall for his valuable comments and discussions of earlier versions of this paper. Thanks are also due to the editors and two anonymous reviewers for their detailed comments.

Bibliography

Aitchison, Jean. 1994. *Words in the mind: An introduction to the mental lexicon*. Oxford: Basil Blackwell.

Drew, Paul & Elizabeth Holt. 1998. Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in Society* 27. 495-522.

Dhillon, Rajdip, Sonali Bhagat, Hannah Carvey & Elizabeth Shriberg. 2004. *Meeting recorder project: Dialog act labeling guide*. ICSI technical report TR-04-002, February 9, 2004. http://www.icsi.berkeley.edu/ftp/pub/speech/papers/MRDA-manual.pdf (accessed 10 September 2009).

Ellson, John, Emden Gansner, Yehuda Koren, Eleftherios Koutsofios, John Mocenigo, Stephen North, Gordon Woodhull, David Dobkin, Vladimir Alexiev, Bruce Lilly, Jeroen Scheerder, Daniel Richard G. & Glen Low. 2005. *Graphviz—Graph visualization software*. http://www.graphviz.org (accessed 10 December 2005).

Galley, Michel, Eric Fosler-Lussier, Hongyan Jing & Kathleen McKeown. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 562–569. http://www.aclweb.org/anthology/P/P03/P03-1071.pdf (accessed 2 May 2006).

Geluykens, Ronald. 1993. Topic introduction in English conversation. *Transactions of the Philological Society* 91(2). 181-214.

Grosz, Barbara & Candace Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics* 12(3). 175–204.

Hirschberg, Julia & Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3). 501–530.

Howe, Mary. 1991. *Topic changes in conversation*. Ph.D thesis, University of Kansas.

Manning, Christopher & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: The MIT Press.

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

Morgan, Nelson, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg & Andreas Stolcke. 2001. The meeting project at ICSI. In *Proceedings of the 1st International Conference on Human Language Technology Research*, 1–7.

http://www.aclweb.org/anthology/H/H01/H01-1051.pdf (accessed 11 July 2004).

Orletti, Franca. 1989. Topic organization in conversation. *International Journal of the Sociology of Languages* 76. 75-85.

Passonneau, Rebecca & Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics* 23(1). 103–139.

Sacks, Harvey. 1992. *Lectures on conversation*. Vol. 2. Oxford: Basil Blackwell.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang & Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, 97–100. http://www.aclweb.org/anthology/W/W04/W04-2319.pdf (accessed 10 July 2004).