Computational Linguistics and Chinese Language Processing
Vol. 27, No. 1, June 2022, pp. 27-52

Topic Development and Boundary Cues in Hakka Conversational Discourse

Shu-Chuan Tseng* and Hsiao-chien Liu+

Abstract

The structure of conversational discourse is context-dependent, and the organization of discourse segments and preferences for signaling discourse boundaries are language-specific characteristics. Participating speakers, speaking scenarios, and communication purposes instantaneously affect the conduct of social interaction and verbal exchanges during a conversation. For example, topic maintenance is sustained by the overt exchange of coherent information, and lexical preferences at the boundaries of related discourse segmentation can help construct the course of topic development. Moreover, form-based discourse units are used to represent the content of spoken utterances and to describe the interaction of speakers in conversations. This study investigated topic-specific Hakka conversations using a top-down two-level discourse segmentation approach to examine the development and production of topics. Typical cues and expressions used to initiate topics and subtopics and their respective discourse functions in the Hakka conversations were analyzed. In the Hakka conversational data, noun phrases were preferred at the topic and subtopic transition boundaries, and complete forms such as clausal constructions were also favored, although the spontaneous speech was expected to be fragmentary in terms of syntactic structure.

Keywords: Conversation, Discourse Units, Topic Development, Boundary Cues, Hakka

1. Introduction

A constituent of a given discourse may be defined as a "contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse context," as proposed by Polanyi (2005: 266). This kind of discourse

* Institute of Linguistics, Academia Sinica
E-mail: tsengsc@gate.sinica.edu.tw
+ College of Hakka Studies, National Central University
E-mail: justcarrie1@gmail.com

segment involves interactive domains, such as discourse genres and speech events, and its segmentation is mainly guided by semantic criteria (e.g., a complete state of affairs and a complete semantic representation), syntax (e.g., clauses and sentence boundaries), and intonation (e.g., pauses and prosodic contours) (Polanyi, 1995). Moreover, discourse segments are indicated by topic shift markers that have been categorized as discourse markers, pragmatic markers, discourse operators, and cue phrases in the literature (van Dijk, 1977b; Grosz & Sidner, 1986; Fraser, 1996; Redeker, 2006). Polanyi (1995) also proposed that discourse operators force segmentation breaks on semantic grounds, as will be shown later in the data from the Hakka conversations. To describe the semantic structure of a conversational discourse, constituent units and their composition/decomposition principles are needed as well as devices to identify boundaries for effective discourse segmentation.

1.1 Discourse Topics

Discourse topics form a coherent discourse that expands on a number of common themes. Van Dijk (1977a: 136) defined discourse topics as "a proposition entailed by the joint set of propositions expressed by the sequence…proposition T is TOPIC of sequence of propositions $\sum = <P_1, P_2,\ldots P_n>$ iff for each $P_i \in \sum$ there is a subsequence $\sum_k$ there is a P_j such that $\sum_k \Rightarrow P_j$ and $T \Rightarrow P_j$": each sequence entails a global proposition P_i, and the global proposition entails a super-global proposition P_j, which is the topic. Giora (1985: 116) also defined "the element relative to which the whole set of propositions (of that segment) is taken to be 'about'"; in other words, a topic should not be derivable from the discourse that occurs before it is introduced. Slightly differently, Geluykens (1993: 118) defined a topic as "information which has a low degree of Recoverability…and which has Persistence": for information to be considered a topic, it must be sustained over a reasonably long stretch of discourse. Stede (2012: 38) gave a clear definition: "A topic as a property of a text segment is characterized by the particular distribution of content words in that segment, and the difference to the distribution in other segments." Todd (2016: 11) combined his own definition with Giora's (1985) "aboutness" and proposed that a topic is determined on the basis of aboutness, connectedness, and relevance. Connectedness is relevant to cohesion and coherence and makes a stretch of language into a meaningful whole. Topics are usually distinguished in terms of their explicitness, with cohesion used for explicit links, or the overt relationship between propositions, and coherence for implicit links, which requires background knowledge or contextual knowledge for interpretation. To identify topic boundaries, cohesion markers, lexis, and coherent concepts are applied. Aboutness is a semantic construct in which all propositions in the discourse are related to a superordinate discourse topic; relevance is concerned with the relationship between a proposition and the one that precedes it, and consistent relevance between propositions makes a discourse coherent.

Asher (2004) applied formal semantic analysis and developed a dynamic theory of

discourse topics by examining four types of contrastive topics: (1) alternation, which incorporates parallel and contrastive notions; (2) narration, which represents two connected discourse segments that appear in a background-foreground event; (3) subordinating and coordinating relationships, which are dependent on the degree of attachment to the antecedents; and (4) summarizers, which are used when there are many discourse segments. Furthermore, Asher and Vieu (2005) noted that in the segmented discourse representation theory, a common topic can be shared by two related constituents (Asher, 1993). Regarding the principles within a topic, the "right frontier constraint" provides attachment points for new information (Webber, 1988: 8). Another proposed principle—continuing discourse patterns—suggested that a coordinating relationship bears the same discourse relationship with a dominating constituent and that the coordinated constituents of a substructure must follow a certain pattern with respect to the dominating constituent (Asher & Vieu, 2005: 595). Subordination and coordination affect topicality in that two constituents are coordinately linked if they contribute to the topic of the larger segment, while they are subordinately linked if one of them is a subtopic (Asher, 1993; van Kuppevelt, 1995a).

Givón (1983) proposed a hierarchical structure that accounted for the preceding discourse context information. In macro structures, thematic paragraphs are larger thematic units that are composed of multi-propositional and chained clauses. Within a thematic paragraph, there are three types of topics—chain initial topics, chain medial topics, and chain final topics—defined by their relative position in a speech flow. A chain initial topic is a "newly introduced, newly changed or newly returned topic" (Givón 1983: 9), and therefore usually has a discontinuous relationship with the preceding context but is potentially persistent in the succeeding context if it introduces an important topic. A chain medial topic is continuous in terms of the preceding context and is persistent, but not maximally so, in the succeeding context. Finally, a chain final topic is continuous in terms of the preceding context but is not persistent in the succeeding context, even if it deals with an important topic. Givón (1983) also defined three quantitative measures—referential distance ("lookback"), potential interference ("ambiguity"), and persistence ("decay")—to describe topic properties in discourse; these measures reflect the degree of topic continuity, topic disruption, and topic persistence, respectively.

Van Kuppevelt (1995a, 1995b) proposed that the topic unit does not always stick to the NEW/OLD principle but appears in different syntactic forms. He further specified that

> the main structure of a bound discourse is determined by one leading discourse topic constituted in one production step at the beginning of the discourse. The development of such a discourse is, with regard to its main structure, from the beginning, bound programmatically by the set topic-constituting questions defining its discourse topic. (van Kuppevelt, 1995a: 139)

The topic hierarchy of a discourse, according to this proposal, contains discourse topics, topics, and subtopics.

1.2 Discourse Segmentation

Discourse segments can be of various sizes and empirically specified by applying operational principles that define their linguistic forms and discourse functions. Within a discourse topic, there is normally a kind of coherent relationship between adjacent lexical chains, which reflects the meaning and function of the discourse. In a conversation, a degree of unity is required to achieve cohesive relationships between sequences of words within a certain stretch of speech, such as reference, ellipsis, substitution, conjunction, and lexical cohesion (Halliday & Hasan, 1976). Coherent relationships between clauses and sentences, such as elaboration, subordination, cause, and exemplification, were also discussed in depth by Mann and Thompson (1988). Morris and Hirst (1991) analyzed lexical cohesion to determine coherence by computing lexical chains in a thesaurus, where thesaural relationships, transitivity of word relations, and distance in sentences allowable between words in a chain were examined. Hoey (2005) used the convergence of overt cohesion and perceptible coherence as the criteria and found that lexical priming and cohesion influenced the comprehensibility of a discourse's organization.

When a topic chain occurs over a succession of several nearby clauses that share a single topic, topic shifts completely direct the discourse text away from the present topic, while topic drifts do not stray far from the present topic. A topic returns if it is mentioned again. Based on Hobbs (1990), three coherence relationships regarding topic drifts have been proposed: if two segments assert propositions with similar or identical properties, then they have a parallel relationship; if a segment serves as a cause for another segment, then this represents an explanation relationship; and if a segment involves the evaluation of comments on a previous topic with no additional new information, then it has an evaluation, or metatalk, relationship.

Cues in topic shifts indicate digressions, but cues in topic drifts may not. Considered in light of Fraser's (1996, 2009) definitions, topic shift cues correspond to topic change markers or digression markers, while topic drift cues can link to other markers of different functions, such as contrast, elaboration, and inference. Todd (2016) considered that there is a continuum from shift and drift to maintenance, rather than these cues forming discrete categories of boundaries. More specifically, drifts are weak boundaries, whereas shifts are strong boundaries. In English, topic boundaries may be marked to signal a shift and attract attention, as in the case of 'Oh, I meant to tell you'. Conversely, 'well' is likely a topic drift marker since it has a much more ambiguous role and is followed by a mix of new and old information. Topic shift markers are used to indicate discourse boundaries, but because of different perspectives and a long history of investigations, they have been given various labels, such as pragmatic connectives (van Dijk, 1977c), discourse particles (Schourup, 1985), discourse markers, transition markers,

discourse operators (Schiffrin, 1987; Fraser, 2006; Redeker, 2006), pragmatic markers (Fraser, 2009), digressive markers (Charolles, 2020), cue phrases (Grosz & Sidner, 1986; Hirschberg & Litman, 1993; Horne *et al*., 2001), clue words (Cohen, 1984), and so on. Quirk (1972) proposed a system of taxonomy that included parallel, inference, summary, detail, reformulation, and contrast markers illustrated by the cue phrases 'in addition', 'as a result', 'in sum', 'in particular', 'in other words', and 'conversely', respectively. Grosz and Sidner (1986), Fraser (1996), and

Table 1. Three proposals for boundary cues

	Functions	Examples
Cue phrases Grosz & Sidner (1986: 198)	attentional change	(push) now, next, that reminds me, and, but (pop to) anyway, but anyway, in any case, now back to (complete) the end, ok, fine, (paragraph break)
	true interruption	I must interrupt, excuse me
	flashback	Oops, I forgot
	digression	By the way, incidentally, speaking of, did you hear about
	satisfaction-precedes	in the first place, first, second, finally, moreover, furthermore
	new dominance	for example, to wit, first, second, and, moreover, furthermore, therefore, finally
Turn-internal discourse segment transitions in spontaneous speech Redeker (2006: 345)	end of segment	okay?, you know, so
	next segment	okay, so, but, now, well, and
	digression, interruption	by the way, you know
	specification, definition	that is, you know, well
	paraphrase	I mean, you know, that is
	explication, clarification	because, you know, I mean
	background information	because, see, well
	comment	you know, I think, I guess
	correction, emendation	oh, or, I mean
	quote	you know, like, well, oh
	return	but (anyway), so, now, well
Pragmatic markers Fraser (1996)	topic change markers	incidentally, speaking of X, parenthetically, by the way, just to update you, that reminds me, before I forget, back to my original point, returning to my point, on a different note
	contrastive markers (denial or contrast)	but, instead, however, all the same, anyway, in any case/rate/event, nevertheless, conversely, despite, even so, regardless, still, that said, though, yet
	elaborative markers (elaboration or refinement)	above all, in other words, what's more, also, alternatively, besides, by the same token, correspondingly, for instance, on top of it all, to cap it all off
	inferential markers (developed based on inference)	after all, so, accordingly, because of this/that, for this/that reason, it can be concluded that, it stands to reason that, of course, then, thus, so

Redeker (2006) all intended to capture the indications of segment transitions, but Fraser's (1996) four discourse markers are much more straightforward. The three proposals for boundary cues are summarized in Table 1.

Boundary cues, as defined in terms of the listed discourse functions in Table 1, are of various linguistic lengths. Words, word sequences, short phrases, and clauses can all serve as boundary cues. In addition to the issue of linguistic units, the conventional use of discourse markers may not correspond precisely to the function of marking topic boundaries because discourse markers are defined as expressing distinctive functions, not as indicating coherent relationships between discourse segments (Harabagiu, 1999). According to the data presented by Das (2014), a majority of topic shift relationships are not explicitly signaled by discourse markers.

This study aimed to investigate the discourse structure of Hakka conversations by applying a top-down two-level annotation schema of discourse segments that form sequences of lexical chains with coherent relationships within sequences and cohesive relationships across sequences. The continuity and maintenance of coherent and cohesive relationships were used as the main judgment criteria for identifying the boundaries of discourse segments. Furthermore, we used form-based units to represent the linguistic content to describe the local environment of the lexical chains. The words and phrases that occurred at the topic and subtopic boundaries were then investigated in the context of discourse segmentation.

2. Method

2.1 Data

This study examined five Taiwan Hakka conversations recorded for the National Digital Language Archive Project. The conversations were produced by six female and four male native Hakka speakers aged between 34 and 60 years old. There are two major variants of Taiwan Hakka: Hailu and Sixian (Hakka Affairs Council, 2017). Our sample included five Hailu and five Sixian speakers. All 10 speakers reported that they were fluent in Hakka and that they spoke Hakka better than Taiwan Mandarin and Taiwanese Southern Min. The pairs of speakers were instructed to talk about a topic of their choice, and each recording session lasted approximately 15 minutes. The content of the conversations was lexically transcribed by the second author whose mother tongue is Sixian Hakka. Word segmentation and part-of-speech tagging were conducted according to the *Dictionary of Frequently Used Taiwan Hakka* published by the Ministry of Education in Taiwan. However, some cases did require discussions and consultations with native speakers and linguists. For example, the negation 無 *mo55* in 美術先生無教个 *mui31 sud2 sin31 sang31 mo55 gau31 gai11* ('what the art teacher did not teach') in one of the Hailu Hakka conversations was listed as an auxiliary word in the dictionary, but

in authentic usage, it can also be a verb, a negative marker, or a negator, depending on the context.

2.2 Annotation of Topics and Subtopics

When the main focus of attention shared by the conversational partners changes, it is considered a topic shift. A conversation segment is labeled "topic" if the concepts and messages exchanged by the interlocutors form a coherent set of interactions. A topic segment in principle presents a high degree of coherence and cohesion in terms of the connectedness, relevance, and aboutness of the information expressed in the conversation. A topic segment can only be annotated if the entire stretch shows a steady and continuing context with topic maintenance. In the framework of lexical cohesion analysis, a straightforward way to identify topic boundaries is to focus on lexical chains (Todd, 2016: 41-43), particularly content word collocations or conceptual associations such as boys-girls, laugh-joke, and bee-honey (Halliday & Hasan, 1976; Todd, 2016). Within each topic, there can be a series of components of interactions that represent different manners of elaborating the topic, and these components are annotated as "subtopics." Subtopics normally appear sequentially but may also overlap and recur, as responses from conversational partners are spontaneous. The identification of subtopic boundaries relies on crucial phrases that establish the relationship of lexical cohesion and that serve as the main clues. For instance, for the topic "family," subtopics such as "places of residence" and "children" may be annotated by the names of places and family members or jobs that are reiterated in consecutive utterances. Depending on the research questions and approaches, there may be different segmentation schemes of conversational discourse, and the annotation of discourse structures is to some degree subjective.

In the current study, we used the two-level discourse segmentation scheme presented above and implemented a procedure to possibly mitigate the level of subjectivity. The five Hakka conversations were first transcribed by a native Hakka speaker and then translated into Mandarin texts that were proofread by three adult native Mandarin speakers. Segmentation into topics and subtopics was conducted by the authors by applying the above principles. Two independent annotators were recruited to evaluate whether the identified boundaries of the topics and subtopics were appropriate for segmenting the conversations. Table 2 lists the results. Both annotators reached agreement in nearly 80% of the topic and subtopic boundaries assigned by the authors. We noticed that in one of the conversations, the rate of disagreement was particularly high, which may have been attributed to a large number of unclear transitions held by the very dominant speaker who produced long topic segments that consisted of several subtopics without a clear boundary. At least one annotator or both annotators did not agree with 20% of the originally segmented boundaries. The location of the boundary was generally agreed by both annotators. Disagreements mostly resulted from deviated judgement about whether a

boundary was a subtopic or a topic. These boundaries were reconsidered and revised by the authors. Eventually, we obtained a final version of discourse segmentation for the Hakka conversations.

Table 2. Topic and subtopic boundary labeling

Boundaries	Hakka Conversations
# of topic boundaries	46
# (%) in agreement	36 (78.26%)
# of subtopic boundaries	261
# (%) in agreement	214 (81.99%)

Below is an excerpt from the data showing a discussion on the topic "language use." The subjects 吾家娘 ('my mother-in-law') and 偌个細人仔 ('my child') were often omitted in the utterances. However, this nominal ellipsis did not hinder the participants' understanding of the speech content, and the topic "language use" clearly remained the focus of successive elaborations until a conclusion was finalized at the end of this discourse segment.

吾家娘乜當希望偌个細人全部講客話	*nga55 ga31 ngiong55 me11 dong53 hi53 mong33 ngai55 gai11 se11 ngin55 cion55 pu33 gong24 hag5 fa11* ('my mother-in-law also expected *my child* to speak Hakka all the time')
佢渡个時節	*gi55 tu33 gai11 shi55 zied5* ('when she [my mother-in-law] took care of [my child]')
全部講客話 hon	*cion55 pu33 gong24 hag5 fa11 hon* ('[my mother-in-law] spoke Hakka all the time')
可是讀書開始	*ko24 shi33 tug2 shu31 koi53 shi24* ('but since [my child] started going to school')
讀幼稚園開始	*tug2 rhiu11 chi55 rhan55 koi53 shi24* ('since [my child] started going to kindergarten')
斯專門講國語啊	*sii53 zhon53 mun55 gong24 gued5 ngi53 a* ('[my child] just spoke Mandarin day and night')
佢成時講國語	*gi55 shang55 shi55 gong24 gued2 ngi31* ('she [my child] spoke Mandarin constantly')
啊國語講啊流流利利	*a gued5 ngi53 gong24 a liu55 liu55 lad3 lad3* ('[*my child*] spoke Mandarin fluently')
講久	*gong24 giu24* ('[my child] had been speaking for a long time')
該客話斯毋記得了 hon	*gai55 hag5 fa11 sii33 m55 gi11 ded5 le31 hon* ('[my child] did not know how to speak Hakka')

2.3 Annotation of Discourse Units

We used a form-based discourse unit (DU) to represent and analyze the discourse structure of the sampled conversations concerning the more information-based discourse segments, topics, and subtopics. In principle, a DU is equivalent to a clause or a sentence in written language. After the main predicate is identified, a DU includes the speech stretch containing the main predicate and the remaining syntactic components, including the subject and the related complements and adjuncts. Some DUs are isolated noun phrases or non-clausal units with no predicates. Non-predicative DUs of this type occur frequently in Japanese and Mandarin Chinese interactional discourse and employ a range of functions, such as referent introduction, identification, and listing (Iwasaki, 1993; Tao, 1996, 2020). In Hakka, verb complexes are often used. To identify DUs in the Hakka data, we referred to the definition of clauses proposed by Thompson and Couper-Kuhlen (2005) and the principles of determining utterance units (Nakajima & Allen, 1993). Please note that DUs are annotated solely based on their constructional form and that both predicative and non-predicative DUs can express complete or incomplete meanings and information. We proposed dividing the DUs into three main types according to their form and meaning: (i) clausal DUs with clear meaning; (ii) non-clausal DUs with clear meaning; and (iii) fragmentary DUs with incomplete meaning. The linguistic content of the Hakka conversations was represented and analyzed in terms of DUs and DU types. Detailed explanations of the DU annotation principles are provided below:

(1) Clausal DUs with clear meaning

a. Clauses delineate complete sentential meanings and satisfy discourse functions, e.g., 佢就渡一個細嬰 *gi55 ciu33 tu33 rhid5 gai11 se11 o53* (‘He only takes care of one baby’) and 偓會講分偓侬仔聽喔 *ngai55 voi33 gong24 bun53 ngai55 lai11 er55 tang11 o* (‘I will tell my son!’). These kinds of DUs often express substantial statements in conversations.

b. DUs with elliptical subject or object NPs that convey a clear and coherent meaning, e.g., 敢還哪看得著客家話 *gam31 han11 nai55 kon55 ded2 do31 hag2 ga24 fa55* (‘Where can we see Hakka language?!’), 聽毋識 *tiang24 m11 siid2* (‘[I] cannot understand’), 來正知个啊 *loi55 zang11 di31 ge55 a* (‘Only when we came did they know that [we are Hakka]’), and 面前就講 *mien55 qien11 qiu55 gong31* (‘[I] talked about it earlier’).

c. Complex DUs that contain focus markers[1] or conditional markers, e.g., 無講若般看人斯做毋得 *mo55 gong24 rhog2 ban53 kon11 ngin55 sii53 zo11 m55 ded5* (‘Not to

[1] The typical Mandarin Chinese focus marker in the cleft constructions 是 *shi* and 是…的 *shi…de* are not exactly the same as 無講, 斯, 係 in Hakka. In this study, they were tentatively considered the focus markers that served the function of emphasis or indications of the upcoming discourse segment.

mention it is not permitted to have a short look at people'), 恁自家愛想愛去哪斯去哪 *an31 qid2 ga24 oi55 xiong31 oi55 hi55 nai55 sii24 hi55 nai55* ('So I go out at will'), and 係無貪就無熟事 *he55 mo11 tam24 qiu11 mo11 sug5 sii55* ('If [you] are not greedy, you will not know [those people]').

(2) Non-clausal DUs with clear meaning

a. This type of DU has no predicate but conveys a clear discourse meaning. These DUs may be used for responses or to introduce a new topic and can take a variety of constructional forms, e.g., 係啊 *he55 a* ('yes'), 正經啊 *ziin55 gin24 a* ('[It is] real'), 吾嫂這兜啊 *nga24 so31 ia31 deu24 a* ('[this situation] applies to people like my sister-in-law'), and 偃苗栗縣个 *ngai11 meu11 lid5 ien55 ge55* ('I [am from] Miaoli County').[2]

b. Predicative adjectives used as part of a verbal complement, e.g., 若若若細人幾大 *ngia24 ngia24 ngia24 se55 ngin11 gi31 tai55* ('how old are your children'), 補助盡高喔 *bu31 cu55 qin55 go24 o* ('the subsidy is high!'), and 恁打爽忒了 *an24 da24 song24 ted5 le53* ('that is a pity'). Thompson and Tao (2010) found that conversational Mandarin speech favors predicate adjectives (80%) over attributive adjectives (20%).

c. Particle DUs that are responsive backchannels, such as 嗯 *n*, 嗯嗯 *n n*, 喔喔 *o o*, and 唉 *ai*, or connective-like junctures that express different speaker attitudes. For instance, the modal particle *hon* in the following example serves as the concluding function and expresses the speaker's intention to obtain approval from the conversational partner in *hon*, e.g., hon...<我們在畫畫>个時節佢會摎人<修改> *hon...wo214 men zai51 hua51 hua51 gai11 shi55 zied5, gi55 voi33 lau31 ngin55 xiu55 gai214* ('hon...he would help students make modifications when we were drawing').

(3) Fragmentary DUs with incomplete meaning

a. DUs that contain noun phrases are used to express the speaker's intention or for communicative functions, such as introducing referents (Iwasaki, 1993; Tao, 1996, 2020). Isolated noun phrases are seldom used to indicate topic shifts and drifts. They may appear together with prepositions or particles, e.g., 對厥印象 *dui11 gia55 rhin11 siong33* ('about the impression of him') and 然後在<那個>[3] 年代 hon 恁仔 *rhan55 heu33 cai33 na31 ge ngien55 toi33 hon an24 ne31* ('then, in that era').

b. Disfluent DUs with no predicates, such as speech repairs or repetitions, e.g., 偃一句偃

[2] 偃苗栗縣个 is considered a non-predicative DU with the nominalization marker 个.

[3] Content appearing in < > was spoken in Mandarin Chinese.

又毋 *ngai11 id2 gi55 ngai11 iu55 m11* ('I cannot even [say] a sentence...') and 僅毋識 *ngai55 m11 siid2* ('I have not been…'). Please note that repairs are not necessarily related to the proposition of the next DU uttered by the same speaker or by the conversational partner, e.g., 嗯恁仔關於講該教育方面个時節…以前个時節你會 (repair)…啊比論講啊你以前無讀著个理想个 *n an24 e31 gon31 rhi55 gong24 gai55 gau11 rhug2 fong31 mien11 gai11 shi55 zied5...rhi31 cien55 gai11 shi55 zied5 ngi55 voi33...a bi24 lun33 gong24 a ngi55 rhi31 cien55 mo55 tug2 do24 gai11 li31 siong24 gai11* ('As for education...before, you would...for example, you have not majored in the ideal subjects...').

2.4 Results

The annotation results of the topics, subtopics, and DUs in the five Hakka conversations are summarized in Table 3. Each conversation covered a different number of distinctive topics that were initiated and discussed by the participants. Please note that the speakers may have restarted a previously discussed topic initiated by themselves or their conversational partners along the course of the conversation. In such cases, we included the occurrences of returning topics in the calculation of topic segment tokens. The number of subtopics per topic was between four and five, but the patterns of speaker interaction and information exchange were in fact individually different in the Hakka conversational data, which will be shown later.

Table 3. Annotation results of the five Hakka conversations

	Con. 1	**Con. 2**	**Con. 3**	**Con. 4**	**Con. 5**
Duration	11 mins	11 mins	14 mins	13 mins	11 mins
# of topics	9	8	11	10	8
# of subtopics	45	53	40	56	31
# of topic segments	12	10	12	13	8
# of subtopic segments	53	54	47	67	35
# of DUs	665	598	878	715	554
# of syllables	3,377	3,320	3,687	3,626	2,810

While the complete coverage of concept exchanges was sustained within a topic, subtopics were operationally more authentic in that they actually formed the continuity of the topic, on the one hand, and connected the consecutive DUs, on the other hand. This also indicated that the ratio of DUs to subtopics was an authentic reflection of topic transitions. The referential distance measurement in Givón's (1983) hierarchical structure proposed that the degree of distancing in topic continuity is 20 clauses in terms of the number of clauses toward the left edge. Additional attempts to measure topic segment length have been proposed in the literature,

for instance, a typical paragraph (Ferret & Grau, 2000), a length of three to five sentences (Hearst, 1993), and a length of approximately 100 words (Dias & Alves, 2005). As shown in Table 3, the degree of speaker activity in the conversations varied, as the number of topics and subtopics initiated by each speaker was considerably different.

The complexity of topics and subtopics to some degree revealed idiolect differences in maintaining topic continuity. Nevertheless, collective commonalities across the Hakka speakers were shown by the number of syllables per DU, which ranged from four to six. Prévot *et al*. (2015) examined DU distributions in French and Taiwan Mandarin conversational data and reported an average DU length of 10.7 syllables for French and 9.6 syllables for Taiwan Mandarin in long speaker turns. Prévot *et al*.'s (2015) study mainly focused on DU components rather than speaker interaction. Compared with Givón's (1983) proposed measure of 20 clauses over a sustained topic, our measurement of DUs per subtopic in Table 4 showed similar results:

Table 4. Annotation results of the five Hakka conversations by speaker

	Con. 1		**Con. 2**		**Con. 3**		**Con. 4**		**Con. 5**	
Speaker gender	F	M	F	M	F1	F2	F	M	F	M
# of syllables	1,158	2,219	1,502	1,818	1,703	1,984	2,709	917	1,289	1,521
# of distinctive topics	5	4	3	5	3	8	7	3	1	7
# of distinctive subtopics	21	24	26	27	18	22	41	15	12	19
# of DUs	294	371	260	338	396	482	471	244	268	286
# of syllables per DU	*3.9*	*6*	*5.8*	*5.4*	*4.3*	*4.1*	*5.8*	*3.8*	*4.8*	*5.3*
# of topic initiations	6	6	4	6	3	9	10	3	1	7
# of subtopic initiations	26	27	26	28	21	26	52	15	14	21
# of DUs per topic segment	49	61.8	65	56.3	132	53.6	47.1	81.3	268	40.9
# of DUs per subtopic segment	*11.3*	*13.7*	*10*	*12*	*18.9*	*18.5*	*9.1*	*16.3*	*19.1*	*13.6*

In addition, more topic and subtopic initiations did not imply more DU production, as shown in Table 4. That is, the degree of active participation in the verbal exchanges was viewed from different perspectives. For instance, in conversation #4, the male speaker clearly initiated fewer new topics, but his active participation was supported by the large number of DUs he produced in taking part in the discussion. On the other hand, compared with his counterpart, he produced shorter DUs that did not deliver complex information as they were mostly responsive. In our two-level discourse segmentation approach, whether topic initiation occurred in response to the previous information delivered by the conversational partner was also an important clue in determining the degree of active participation.

3. Discourse Organization in Hakka Conversations

Understanding conversational discourse requires a structural description of how the discourse is organized. Therefore, it is necessary to have a system of segmentation units whose relationships can be empirically defined. In this study, we annotated three types of units—topics, subtopics, and DUs. Topics and subtopics were identified from a top-down perspective, in which the information content was the principal criterion. The DUs were mainly identified according to their constructional forms. In particular, predicates were used to categorize the types of DUs.

3.1 Conversational Discourse Descriptions

Todd (2016: 172) mentioned that topic development is divided into three main types—maintenance, drift, and shift—that can be further categorized into subtypes, such as major and minor shifts. In our annotation of topics and subtopics, we took these main types into consideration to describe the interplay of information exchanges and transactions in the discourse organization of the Hakka conversations. Figure 1 shows the hierarchical structure of the topics and subtopics represented by the DUs extracted from the data. The DUs produced by *Speaker A* are underlined in the transcript, and speech content uttered at topic and subtopic boundaries are in boldface. The interaction of the speakers and the speech production patterns is illustrated in terms of this representational format. The identification of topic shift and drift was content-based, while the DUs were defined according to the placement and scope of the predicates.

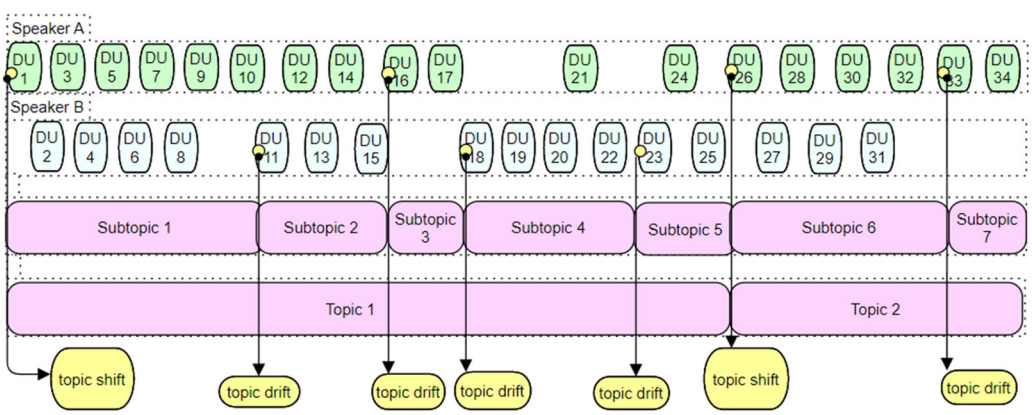

Figure 1. Topic development in the conversations

Topic 1: Speaking Hakka[4]

Subtopic 1: It's difficult for Southern Min people to learn Hakka

DU1: **你講故所講學老人愛學偃兜客也當難** *ngi11 gong31 gu55 so31 gong31 hog5 lo31 ngin11 oi55 hog5 en34 li55 hag2 ia24 dong24 nan11* ('So [it's] also very difficult for Southern Min natives to learn Hakka')

DU2: 嗯嗯… *n n*… ('um um…')

DU3: en24 li55 客人愛學學老較 goi24 啦 *en24 li55 hag2 ngin11 oi55 hog5 hog5 lo31 ka55 goi24 la* ('It is easier for us Hakka people to learn Southern Min')[5]

DU4: 嗯嗯… *n n*… ('um um…')

DU5: 因為電視台不時會做啊 *in24 vi55 tien55 sii55 toi11 bud2 sii11 voi55 zo55 a* ('Because there are often [Southern Min] TV programs')

DU6: <對啊> *dui4 a* ('correct')

DU7: <連續劇>唱歌仔就唱學老 *lian2 xu4 ju4 cong55 go24 qiu55 cong55 hog5 lo11* ('Those who sing in serials sing Southern Min')

DU8: 係係 *he55 he55* ('yes yes')

DU9: 故所偃兜客家人當會去講學老啊 *gu55 so31 ngai11 deu24 hag2 ga24 ngin11 dong24 voi55 hi55 gong31 hog5 lo31 a* ('So we, Hakka people, are good at speaking Southern Min')

DU10: 學老人會講客家話就當難啊 *hog5 lo31 ngin11 voi55 gong31 hag2 ga24 fa55 qiu55 dong24 nan11 a* ('It's difficult for Southern Min people to speak Hakka!')

Subtopic 2: You are good at speaking Southern Min

DU11: **該你學老乜當厲害喔** *ge55 ngi11 hog5 lo31 me55 dong24 li55 hoi5 o* ('You are good at Southern Min')

DU12: 會講啦 *voi55 gong31 la* ('[I] can speak [it]')

DU13: 會講啦係囉 *voi55 gong31 la he55 lo* ('[You] can speak [it]')

DU14: 講講毋會當滑溜啦 *gong31 gong31 m11 voi55 dong24 vad5 liu55 la* ('[I] cannot speak it very fluently')

DU15: 嗯嗯… *n n*… ('um um…')

---

[4] Topic 1 "Speaking Hakka" was the common property shared by Subtopics 1 to 5.

[5] Subtopic 1 "It's difficult for Southern Min people to learn Hakka" described the unfair situation that it is easier for Hakka natives to learn Southern Min but not that easy for Southern Min natives to learn Hakka. What the speakers focused on was comparing the scenarios, so it was not appropriate to classify DU3 and DU4 into another subtopic different from Subtopic 1. This part was approved by the two independent annotators in our boundary segmentation evaluation experiment.

Subtopic 3: Hakka people are poor

DU16: **故所講 hon** *gu55 so31 gong31 hon* ('So')

DU17: 客家人盡衰過啊 *hag2 ga24 ngin11 qin55 coi24 go55 a* ('Hakka people are very poor')

Subtopic 4: Zhunan Hakka group

DU18: **毋係啦** *m11 he55 la* ('No, it is not true')

DU19: 一方面 *hon id2 fong24 mien55 hon* ('On the one hand')

DU20: 你像偃蹄竹南 *ngi11 qiong55 ngai11 dai55 zug2 nan11* ('For example, I live in Zhunan')

DU21: 假使乜學老人 *ga31 sii31 me55 hog5 lo31 ngin11* ('There are supposedly many Southern Min people')

DU22: 著 *cog5* ('Yes')

Subtopic 5: Hakka people are afraid to speak Hakka

DU23: **<百分之六十>个客家人** *bai3 feng1 zhi1 liu4 shi2 ge55 hag2 ga24 ngin11* ('About sixty percent of the Hakka people')

DU24: 毋敢講客 *m11 gam31 gong31 hag2* ('are afraid to speak Hakka')

DU25: 係啊 *he55 a* ('Yes, it is true')

Topic 2: Policy related to Hakka people

Subtopic 1: Minister of the Council of Hakka Affairs

DU26: **這擺斯好得<葉菊蘭>hon** *ia31 bai31 sii11 ho31 ded2 ye4 ju2 lan2 hon* ('It is good to have Yeh Chu-lan this time')

DU27: 係 *he55* ('Yes')

DU28: 恁樣 *an31 ngiong11* ('What')

DU29: <客家會主委> *ke4 jia1 huei4 zhu2 wei3* ('Minister of the Council of Hakka Affairs')

DU30: 佢毋係<客委會主委> *gi11 m11 he55 ke4 wei3 huei4 zhu2 wei3* ('She is not the minister of the Hakka Affairs Council')

DU31: 佢已早盡早係啦 *gi11 i31 zo31 qin55 zo31 he55 la* ('She was once the minister')

DU32: 係 *he55* ('Yes')

Subtopic 2: Democratic Progressive Party (DPP)

DU33: **講下擺講** *gong31 ha55 bai31 gong31* ('Speaking of')

DU34: **恁樣民進黨愛仰仔講 hon** *an31 ngiong11 min11 jin55 dong31 oi55 ngiong31 e31 gong31 hon* ('How to describe the DPP')

Understanding a conversation, for both humans and automatic systems, is to steadily obtain new and recurrent patterns of information about social interaction, linguistic content, and speaker intention. Judgments and annotations that refer to previously uttered conversational speech data are in fact indirect evidence of language planning processes inferred from the shared information between conversational partners. Nevertheless, a representational model, as we have suggested, provides a hierarchy of discourse components within a sequence of verbal interactions, and DU-initial cue words can be used to tackle the rhetorical relationships between the uttered propositions for linguistic research as well as to heuristically identify topic segmentation boundaries to enhance semantic understanding in natural language processing.

3.2 Interaction and Linguistic Patterns

Within the interaction that occurs during a conversation, new and recurrent topics and subtopics may be initiated by non-responsive or responsive actions. Tables 5 and 6 summarize the results of the DUs in terms of the three DU types, clausal, non-clausal, and fragmentary, divided into two interaction categories, non-responsive and responsive. Please note that if the discourse meaning of a DU was incomplete, that is, the speech content could not be clearly interpreted and specified, it was classified as a fragmentary meaning. The DUs whose discourse meaning could be clearly identified were divided into clausal and non-clausal meanings.

Table 5. DU types used for topic initiation

		Con. 1		**Con. 2**		**Con. 3**		**Con. 4**		**Con. 5**	
		F	M	F	M	F1	F2	F	M	F	M
Meaning	**Form**	**Non-responsive (37)**									
clear	*clausal*	1	2	2	4	1	7	4			3
clear	*non-clausal*					1	1				
incomplete	*fragmentary*	1	2	1		1		1		1	4
Speaker's DU proportion		33%	67%	75%	67%	100%	89%	50%		100%	100%
Meaning	**Form**	**Responsive (18)**									
clear	*clausal*	4					1	3	1		
clear	*non-clausal*			1	2				1		
incomplete	*fragmentary*		2					2	1		
Speaker's DU proportion		67%	33%	25%	33%		11%	50%	100%		

Table 6. DU types used for subtopic initiation

		Con. 1		Con. 2		Con. 3		Con. 4		Con. 5	
		F	M	F	M	F1	F2	F	M	F	M
Meaning	**Form**	**Non-responsive (158)**									
clear	*clausal*	9	11	8	9	8	17	18		7	13
clear	*non-clausal*	6		1	1	3	4	1		1	
incomplete	*fragmentary*	1	10	3	2	5	3	6	1	5	5
Speaker's DU proportion		60%	78%	48%	41%	76%	92%	48%	7%	93%	86%
Meaning	**Form**	**Responsive (98)**									
clear	*clausal*	10	2	7	10	3	2	23	11	1	3
clear	*non-clausal*		1	5	5	1		2	2		
incomplete	*fragmentary*		3	1	2	1		2	1		
Speaker's DU proportion		40%	22%	52%	59%	24%	8%	52%	93%	7%	14%

The overall results showed that clausal DUs were the most frequent forms in the conversational data (32/55 for topics, 170/256 for subtopics), although spontaneous conversational speech was expected to be fragmentary in terms of syntactic structure. The results support the notion that complete and coherent forms are favored in producing a locus of interaction and projecting the speaker's actions (Thompson & Couper-Kuhlen, 2005). Fragmentary DUs that had a less clear discourse meaning were actually in the minority, suggesting that when the speakers intended to initiate a new topic or subtopic, the action was to some degree already planned before the topic- or subtopic-initial DUs were produced. Our analysis also showed that the proportion of responsive DUs used for topic shifts (33%) was slightly smaller than that used for topic drifts (38%). This implied that even though we segmented the conversations into coherent topics and subtopics with different degrees of topic continuity, it was nevertheless essential for speakers to provide reactions that responded to their conversational partners' previous verbal actions.

It is noteworthy that backchannels normally refer to short utterances that are produced by the non-primary speaker or the listener when the front channel is occupied by the primary speaker, according to Yngve (1970). In the Hakka data, backchannels such as particles and short replies (e.g., 嗯 *n* and 喔 *o*) also occurred at the topic and subtopic boundaries, and they were used to initiate a new discourse segment while responding to their conversational partner at the same time. The categorization of responsive versus non-responsive DUs contributed to an understanding of conversation interaction. For instance, the male speaker in conversation #4 started 93% of the subtopics by responding to his conversational partner's previous reaction, whereas the other speakers in the data mostly initiated subtopics without directly reacting to

their partners. The distribution of responsive DUs between the two speakers in the conversation was regarded as an indicator that represented the speaker's interaction pattern for participation behavior in a cooperative context. Currently, we are not yet in the position to claim that this is an effective indicator. Nevertheless, we have shown that in addition to linguistic patterns, the analysis of boundary DU types provided insight into the social interaction in the conversational discourses.

4. Initiation Cues in Hakka

Cue phrases in a written or spoken discourse in general refer to connectives and discourse markers that designate relevant positions for discourse segmentation and interpretation. However, to what extent discourse segments of a broader scope, such as topics and subtopics, are signaled by boundary cues with a similar function as cue phrases has not been thoroughly studied. In this study, we attempted to obtain an overview of boundary cues that were recurrently used to initiate topics and subtopics based on the annotation results of the topic and subtopic boundaries.

4.1 General Types of Boundary Cues

Cue phrases are considered pivots that deliver, change, and return linguistic messages. In particular, they are used to signal the speaker's intention for language planning and to attract the listener's attention, given that the coherence relationships between the already-expressed and the to-be-expressed propositions are intact. They may also be regarded as a type of discourse marker. Different from the conventional notion of cue phrases, strong and weak topic boundaries can be signaled by linguistic forms of different lengths, such as words, phrases, and clauses. For instance, 佢摎你講 le *ngai11 lau24 ng11 gong31 le* ('let me tell you') and 佢會講 *ngai11 voi55 gong31* ('I would say'), produced at the topic and subtopic boundaries in the data had a function similar to that of cue phrases. We tentatively included the whole expression 佢摎你講 and 佢會講 in the category of "empirical marker."

Table 7 lists the types of boundary cues that were used to mark topic and subtopic transitions, in which the lexical chain of a new discourse segment with either a broad (topic) or a narrow (subtopic) scope occurred. Some of the boundary cues were language-specific characteristics in Hakka and thus worthy of further investigation. In general, there was no significant difference in the distributions in terms of topics and subtopics. Noun phrases were mostly preferred for initiating topics and subtopics in Hakka, followed by connectives. The particles identified in Table 7 were not used as backchannels but instead served the discourse function of preparing the listeners for the upcoming discourse segments by signaling new information that could change the topics or subtopics. Future studies should further investigate the intonation patterns of backchannel particles and initiation cue particles to elaborate the

relationship between discourse functions and phonetic forms (Hirschberg & Litman, 1993). Empirical markers and negation markers were also identified as boundary cue types, but they did not appear as often as noun phrases, connectives, and particles.

Table 7. Occurrences of boundary cue types

Types	Topics	Subtopics
Noun phrase	16 (29%)	96 (38%)
Connective	21 (38%)	83 (32%)
Particle	13 (24%)	42 (16%)
Empirical marker	3 (5%)	26 (10%)
Negation marker	2 (4%)	9 (4%)
Total	**55 (100%)**	**256 (100%)**

4.2 Analysis of Boundary Cues

We depicted the topic development and speaker interaction in the conversations by topics, subtopics, and DUs. Utilizing this representational format, we examined initiation cues at strong and weak discourse boundaries. Different from the typical English cue phrases in Table 1, Hakka conversations do not exhibit a strong tendency to use specific groups of cue phrases that are in turn used to mark the locations of topic transitions. To gain an overview of the discourse functions of initiation cues in Hakka conversations, we conducted a pilot study. Referring to previous studies on connectives and cue phrases, we attempted to clarify the discourse functions of the boundary cues included in the results presented in Table 7. We did not implement any a priori restrictions on the length of the linguistic units, such as words or phrases, but mainly referred to recurrent patterns to specify their discourse functions. The results summarized in Table 8 are exclusively valid for our data. Herewith, we hope to provide a preliminary system of initiation boundary cues in Hakka conversations that can be further specified in more detail as well as more types of speech data.

Initiating a new discourse segment by specifying objects or qualities is a common practice in Hakka conversations. This may well explain why many noun phrases are used for topic and subtopic initiation, in addition to connectives. Most boundary cues are used for both topics and subtopics; however, in some cases that express concrete specifications of object descriptions and qualities, they do not occur at topic transition positions but are exclusively used at subtopic boundaries. We observed that boundary cues had the function of attracting the listener's attention for a topic transition. When combined with the use of lexically explicit discourse markers, that is, with a clear correspondence of function and meaning, the transition of topics and subtopics was successful and proceeded fluently within the conversations.

Table 8. Boundary cues in Hakka

Types	Functions	Boundaries	Typical Boundary Cues[6]
Noun phrase	Identifying time	Topic	頭擺 *teu11 bai31* (‘before’) 這下 *lia31 ha55* (‘now’)
		Subtopic	頭擺/頭過 *teu11 bai31*/*teu55 go11* (‘before’) 這下 *lia31 ha55* (‘now’) 該下 *ge55 ha55* (‘that time’)
	Identifying objects	Topic	佢个 *gi11 ge55* (‘his’)
		Subtopic	這兜 *ia31 deu24* (‘these’) 該路 *ge55 lu55* (‘that road’)
	Identifying places	Topic	該位 *gai55 vui33* (‘that place’) 你這 *ng11 lia31* (‘your place’)
		Subtopic	佢个這位 *ga11 ge55 ia31 vi55* (‘his place’)
	Identifying persons	Topic	你 *ng11* (‘you’) 俚 *ngai11* (‘me’) 佢 *gi11* (‘he’) 俚等 *ngai11 den31* (‘we’) 該 *ge55* (‘that’)
		Subtopic	你 *ng11* (‘you’) 俚 *ngai11* (‘me’) 佢 *gi11* (‘he’) 俚等 *ngai11 den31* (‘we’) 該 *ge55* (‘that’) 這 *ia31* (‘this’) 這兜 *ia31 deu24* (‘these’)
Connective	Explication, clarification, inference	Topic	所以 *so31 i24* (‘so’) 因為 *in 24 vi55* (‘because’)
		Subtopic	所以 *so31 i24* (‘so’) 故所 *gu55 so31* (‘so’) 因為 *in 24 vi55* (‘because’)
	Contrast	Topic	毋過 *m11 go55* (‘but’)
		Subtopic	毋過 *m11 go55* (‘but’) 可是 *ko31 sii55* (‘but’)
	Topic change	Subtopic	那 *na55* (‘that’)
		Topic & subtopic	該 *ge55* (‘that’)
	Sequence	Topic	過忒 *go55 ted2* (‘then’)
		Subtopic	過了 *go55 e31* (‘then’) 過 *go55* (‘then’) 然後 *ien11 heu55* (‘then’)
	Addition	Topic	還 *han11* (‘also’) 還有 *han11 iu24* (‘in addition’)
		Subtopic	還 *han11* (‘also’)
	Concession, elaboration	Topic	其實 *ki11 siid5* (‘in fact’) 恁多 *an31 do24* (‘so many’)
		Subtopic	其實 *ki11 siid5* (‘in fact’) 假使 *ga31 sii31* (‘if’) 恁 *an31* (‘such’) 敢還 *gam31 han11* (‘could it still be said that...’)
	Relation	Subtopic	像 *qiong55* (‘like’)
Particle	Elaboration, clarification, reaffirmation	Topic & subtopic	*hon* 唉 *ai* 唉喔 *ai o* 啊 *a* 喔 *o* 嗯 *n* 嘖 *jid* 詒 *e*

[6] The boundary cues were transcribed based on the Sixian Hakka dialect.

Empirical marker	Topic change	Topic	故所講 *gu55 so31 gong31* (‘you say, so say’) 偓摎你講 *le ngai11 lau24 ng11 gong31 le* (‘let me tell you’) 偓會講 *ngai11 voi55 gong31* (‘I would say’) 就講 *qiu55 gong31* (‘just speaking’)
		Subtopic	偓就講 *ngai11 qiu55 gong31* (‘I just say’) 下擺講 *ha55 bai31 gong31* (‘sometimes speaking of’) 就講 *qiu55 gong31* (‘just speaking’) 你看 *ngi11 kon55* (‘you see’) 講 *gong31* (‘saying’)
Negative	Negation	Topic & subtopic	毋係 *m11 he55* (‘not’) 無 *mo11* (‘no’)

5. Discussion

The semantic structure of conversational discourse needs to account for the macrostructure of the discourse and the social interaction between the conversational participants (van Dijk, 1977b). References to a given discourse referent may constantly change along the course of a conversation due to spontaneous language planning and speaker reactions. Therefore, sentence-level distinctions of topics and comments may not explicitly or effectively apply to conversational discourse descriptions (van Dijk, 1980; Asher, 2004). Our analysis of Hakka conversational data revealed that linguistic forms represented in terms of predicate-based DUs were useful in presenting the quantity and the quality of content across the subtopics. Subtopics may be more closely connected with the constructional form than a broader sense of discourse segments, such as topics defined by lexical cohesion and coherence relationships (Halliday & Hasan, 1976; Morris & Hirst, 1991; Harabagiu, 1999). Likewise, the distinction between responsive and non-responsive action types, which is important in interpreting the social interaction of participating speakers, is also more conclusive at the level of subtopics rather than topics (Hobbs, 1990; van Kuppevelt, 1995a, 1995b).

Mann and Thompson (1988) proposed rhetorical relations of propositions. If we had intended to apply the rhetorical relationship approach to decompose the content of the conversational discourses into a structured organization, we would have needed to be equipped with a sentence-comparable unit. We adopted the concept of DUs (Grosz & Sidner, 1986; Polanyi, 1995, 2005; Tao, 1996; Prévot *et al*., 2015) to construct elementary units with which higher-level discourse segments could be built. Our results showed that DUs were effective means to link boundary cue types through discourse organization. For topic and subtopic initiation, clausal DUs are preferred (Thompson & Couper-Kuhlen, 2005). Discourse is organized based on coherence relationships that construct the “aboutness” of linguistic segments. Specifically, clauses were proven to be interactionally accessible units in our Hakka data, and our results in Tables 5 and 6 support the notion that the Hakka speakers in this study preferred clausal constructions as a linguistic strategy for topic transitions. In addition, the

discourse meaning of the DUs at the topic and subtopic boundaries also tended to be complete, suggesting that the speakers may have already completed their language planning before they produced upcoming topics.

Givón (1983) and van Kuppevelt (1995a) both proposed a hierarchical structure of discourse topics, with Givón emphasizing a horizontal relationship between the preceding discourse contexts and the current one, and Van Kuppevelt proposing that a discourse is decomposed into discourse topics, topics, and subtopics. According to Givón (1983: 12), when "lookback" is employed as a measure of topic continuity, the upper limit is 20 clauses from the previous occurrence, depending on what "the speaker makes about topic-availability to the hearer, involving the transition from 'availability' or 'identifiability' to the more neutral 'continuity'." It is empirically practical to rely on the principle of continuity, rather than that of discontinuity or disruption, when carrying out the task of discourse segmentation. We proposed a similar, two-level approach for describing discourse organization and topic development in Hakka conversations. The topics and subtopics were mainly identified according to topic continuity and coherence relationships. However, to achieve an understanding of the interaction within a conversation, it was necessary not only to examine the components and their relationships but also to reveal their discourse functions and the social action of the speakers. Our approach preliminarily proved useful in accounting for the linguistic characteristics of the use of DU types and boundary cues. To study speakers' social interaction in interactive conversational speech also requires cognitive accounts that consider the intention and attention status of the conversational partners. That is, a mechanism that provides a link between cognitive states and the corresponding language production is needed (Stede, 2012; Todd, 2016). DUs, as proposed in our approach, may serve as an adequate unit for this purpose.

In the current study on Hakka conversations, we started with the discourse segmentation of the topics and subtopics by applying lexical cohesion analysis. The DUs were identified by referring to the availability of predicates, subjects, and objects according to surface structures. Following this line of data processing, we further specified the discourse functions of the boundary cues to initiate the topics and subtopics. The topic boundary cues were not limited to the specific word category of "cue phrases" but included word sequences that were recurrently used for topic and subtopic transitions. Not only were connectives and particles commonly used in spoken discourse, noun phrases that specified physical objects and qualitative properties were also preferred at the boundaries across topics and subtopics in the Hakka conversations (Tao, 2020).

6. Conclusion

Shared knowledge and semantic coherence are required for the successful execution of conversations. Dynamic changes in coherence relationships in broad and narrow senses

construct the building blocks of conversational discourse. We pointed out that predicate-based clausal accounts of DUs are an operable means of bridging information-based topic segmentation and form-based lexical processing. More studies are needed to account for linguistic properties that are directly related to social behavior, such as an effective means of making discourse segments coreferential to one another, including the use of words, sounds, prosody, and non-verbal elements. We proposed a hierarchical schema to analyze the macrostructure of conversations consisting of topics and subtopics represented in terms of DUs. Systems with more levels of discourse segments are also possible, but according to our results, the subtopics were robust units with which interactive patterns of the conversations were reflected and described. Further empirical studies examining the relationship between discourse segments, initiation cues, and phonetic forms are needed. To meet the goal of understanding and representing a conversational discourse for humans and automatic systems, it is necessary to engage in interdisciplinary collaborations to develop applicable data-driven methodologies for the automatic extraction of coherent and cohesive relationships between topics, as well as sensible mechanisms of cognitive devices that represent the intention and attention states of conversational partners.

References

Asher, N. (1993). *Reference to abstract objects in discourse*. Kluwer Academic Press.

Asher, N. (2004). Discourse topics. *Theoretical Linguistics*, *30*(2-3), 163-201. https://doi.org/10.1515/thli.2004.30.2-3.163

Asher, N., & Vieu, L. (2005). Subordinating and coordinating discourse relations. *Lingua*, *115*(4), 591-610. https://doi.org/10.1016/j.lingua.2003.09.017

Charolles, M. (2020). Discourse topics and digressive markers. *Journal of Pragmatics*, *161*, 57-77. https://doi.org/10.1016/j.pragma.2020.01.005

Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting on Association for Computational Linguistics*, 251-258. https://doi.org/10.3115/980491.980546

Das, D. (2014). *Signalling of coherence relations in discourse* (Doctoral dissertation). Simon Fraser University.

Dias, G., & Alves, E. (2005). Discovering topic boundaries for text summarization based on word co-occurrence. In N. Nicolas et al. (Eds.), *Recent advances in natural language processing IV: Selected papers from RANLP 2005*, 187-191.

Ferret, O., & Grau, B. (2000). A topic segmentation of texts based on semantic domains. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000)*, 426-430.

Fraser, B. (1996). Pragmatic markers. *Pragmatics*, *6*(2), 167-190. https://doi.org/10.1075/prag.6.2.03fra

Fraser, B. (2006). Towards a theory of discourse markers. In K. Fischer (Ed.), *Approaches to discourse particles* (pp.189-204). Elsevier.

Fraser, B. (2009). Topic orientation markers. *Journal of Pragmatics*, *41*(5), 892-898. https://doi.org/10.1016/j.pragma.2008.08.006

Geluykens, R. (1993). Topic introduction in English conversation. *Transactions of the Philological Society*, *91*(2), 181-214. https://doi.org/10.1111/j.1467-968X.1993.tb01068.x

Giora, R. (1985). A text-based analysis of nonnarrative texts. *Theoretical Linguistics*, *12*(2-3), 115–136. https://doi.org/10.1515/thli.1985.12.s1.115

Givón, T. (1983). Topic continuity in discourse: An introduction. In T. Givón (Ed.), *Topic continuity in discourse* (pp. 1-42). John Benjamins.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175-204.

Hakka Affairs Council (客家委員會). (2017). *Survey on national Hakka population and language basic data*, 2017 (105 年度全國客家人口暨語言基礎資料調查研究), Report of Hakka Affairs Council (客家委員會研究報告).

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Harabagiu, S. M. (1999). From lexical cohesion to textual coherence: A data driven perspective. *Journal of Pattern Recognition and Artificial Intelligence*, *13*(2), 247-265. https://doi.org/10.1142/S0218001499000148

Hearst, M. A. (1993). *Text tiling: A quantitative approach to discourse segmentation*. Technical Report Sequoia 93/24. University of California, Berkeley.

Hirschberg, J., & Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, *19*(3), 501-530.

Hobbs, J. R. (1990). Topic drift. In B. Dorval (Ed.), *Conversational organization and its development* (pp. 3-22). Ablex.

Hoey, M. (2005). *Lexical priming*. Routledge.

Horne, M., Hansson, P., Bruce, G., Frid, J., & Filipsson, M. (2001). Cue words and the topic structure of spoken discourse: The case of Swedish men 'but'. *Journal of Pragmatics*, *33*(7), 1061-1081. https://doi.org/10.1016/S0378-2166(00)00044-8

Iwasaki, S. (1993). *Subjectivity in grammar and discourse: Theoretical considerations and a case study of Japanese spoken discourse*. John Benjamins.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243-281. https://doi.org/10.1515/text.1.1988.8.3.243

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*(1), 21-48.

Nakajima, S., & Allen, J. F. (1993). *A study on prosody and discourse structure in cooperative dialogues*. 1993 Technical Report. Department of Computer Science, University of Rochester, NY.

Polanyi, L. (1995). *The linguistic structure of discourse*. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.5029&rep=rep1&type=pdf

Polanyi, L. (2005). The linguistic structure of discourse. In D. Schiffrin et al. (Eds.), *The handbook of discourse analysis* (pp. 265-281). Blackwell Publishers.

Prévot, L., Tseng, S.-C., Peshkov, K., & Chen, A. C.-H. (2015). Processing units in conversation: A comparative study of French and Mandarin data. *Language and Linguistics*, *16*(1), 69-92. https://doi.org/10.1177/1606822X14556605

Quirk, R. (1972). *A grammar of contemporary English*. Longman.

Redeker, G. (2006). Discourse markers as attentional cues at discourse transitions. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 339-358). Elsevier.

Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press.

Schourup, L. C. (1985). *Common discourse particles in English conversation*. Routledge.

Stede, M. (2012). *Discourse processing*. Morgan and Claypool Publishers.

Tao, H. (1996). *Units in Mandarin conversation: Prosody, discourse, and grammar*. John Benjamins.

Tao, H. (2020). NP clustering in Mandarin conversational interaction. In S. A. Thompson et al. (Eds.), *The "noun phrase" across languages: An emergent unit in interaction* [Typological Studies in Language 128] (pp. 271-314). John Benjamins.

Thompson, S. A., & Couper-Kuhlen, E. (2005). The clause as a locus of grammar and interaction. *Discourse Studies*, *7*(4-5), 481-505. https://doi.org/10.1177/1461445605054403

Thompson, S. A., & Tao, H. (2010). Conversation, grammar, and fixedness: Adjectives in Mandarin revisited. *Chinese Language and Discourse*, *1*(1), 3-30. https://doi.org/10.1075/cld.1.1.01tho

Todd, R. W. (2016). *Discourse topics*. John Benjamins Publishing Company.

Van Dijk, T. A. (1977a). *Text and context*. London: Longman.

Van Dijk, T. A. (1977b). Sentence topic and discourse topic. *Papers in Slavic Philology (PSP)*, *1*, 49-61.

Van Dijk, T. A. (1977c). Pragmatic connectives. *Interlanguage Studies Bulletin*, *2*(2), 77–93.

Van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse , Interaction, and Cognition*. L. Erlbaum Associates.

Van Kuppevelt, J. (1995a). Discourse structure, topicality and questioning. *Journal of Linguistics*, *31*(1), 109-147. https://doi.org/10.1017/S002222670000058X

Van Kuppevelt, J. (1995b). Main structure and side structure in discourse. *Linguistics*, *33*(4), 809-833. https://doi.org/10.1515/ling.1995.33.4.809

Webber, B. L. (1988). *discourse deixis and discourse processing* (Technical Report MS-CIS-86-74). Linc Lab 42. Department of Computer and Information Science. University of Pennsylvania.

Yngve, V. (1970). On getting a word in edgewise. In *Papers of the Sixth Regional Meeting of Chicago Linguistic Society*, 567-577.