Annotation of anaphoric relations and topic continuity in Japanese conversation

Natsuko Nakagawa[*†] **Yasuharu Den**[‡]

[*]Department of Linguistics, State University of New York at Buffalo
Buffalo, NY 14260 USA
[†]Graduate School of Human and Environmental Studies, Kyoto University
Yoshida-nihonmatsu-cho, Sakyo-ku, Kyoto 606-8501 Japan
[‡]Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522 Japan
[*†]nakagawanatuko@gmail.com [‡]den@cogsci.l.chiba-u.ac.jp

Abstract

This paper proposes a basic scheme for annotating anaphoric relations in Japanese conversations. More specifically, we propose methods of (i) dividing discourse segments into meaningful units, (ii) identifying zero pronouns and other overt anaphors, (iii) classifying zero pronouns, and (iv) identifying anaphoric relations. We discuss various kinds of problems involved in the annotation mainly caused by on-line processing of discourse and/or interactions between the participants. These problems do not arise in annotating written languages. This paper also proposes a method to compute topic continuity based on anaphoric relations. The topic continuity involves the information status of the noun in question (given, accessible, and new) and persistence (whether the noun is mentioned multiple times or not). We show that the topic continuity correlates with short-utterance units, which are determined prosodically through the previous annotations; nouns of high topic continuity tend to be prosodically separated from the predicates. This result indicates the validity of our annotations of anaphoric relations and topic continuity and the usefulness for further studies on discourse and interaction.

Keywords: anaphoric relations, topic continuity, Japanese conversation corpus

1. Introduction

This paper proposes a basic scheme for annotating anaphoric relations in Japanese conversations and discusses various kinds of problems involved in the annotation mainly caused by on-line processing of discourse and/or interactions between the participants. These problems do not arise in annotating written languages. This paper also proposes a method to compute topic continuity based on anaphoric relations. The topic continuity involves the information status of the noun in question (given, accessible, and new) and persistence (whether the noun is mentioned multiple times or not). We show that the topic continuity correlates with short-utterance units, which are determined prosodically through the previous annotations. This result indicates the validity of our annotations of anaphoric relations and topic continuity and the usefulness for further studies on discourse and interaction.

We focus on issues specific to spoken Japanese in this paper since more general problems associated with anaphora in Japanese are discussed in the literature in the context of annotating written Japanese (Nakaiwa et al., 1995; Hashida, 2005; Iida et al., 2007; Sasano et al., 2008, inter alia).

The outline of this paper is the following: in §2., we discuss the purpose of our study and difference between the previous studies and our work in terms of the goals, interests, and methodologies. In §3., we elaborate the procedure of the annotation. In §4., we investigate the correlation between our annotation and chunks of information in utterance production. In §5., we discuss the remaining issues. In §6., we briefly summarize the current study and suggest future studies.

2. Background

2.1. Purpose of the study

In functional linguistics, sociolinguistics, and various kinds of communication studies, many researchers have been interested in discourse structures, especially in spoken languages, and discuss the relationships between discourse structures and other features such as types of nouns (e.g., reduced (pronoun) vs. non-reduced (full NP)), prosodic characteristics (e.g., whether there is a stress in a given word), and gestures (Chafe, 1994; Givón, 1983; McNeill et al., 2001, inter alia). They are interested in how a topic in discourse is managed and maintained by the participants during the flow of discourse and how the occurrence of topics correlates with the features above.

Schemes for annotating anaphoric relations in spoken languages are useful especially for this kind of research because the annotation makes it possible to keep track of continuous topics which are mentioned several times during the flow of discourse. Since there are very few corpora with anaphoric relations in spoken Japanese, we annotated anaphoric relations in a corpus of casual conversations in Japanese and will discuss various issues related to the annotation.

2.2. Terminology

We use the term *discourse element* (DE) to refer to all NPs occurring in a discourse, including both overt pronouns and zero pronouns. A DE and another DE which refer to the same concept or entity are in *anaphoric relation*. The DE which is in anaphoric relation with another DE and is followed by (an)other DE(s) is the *antecedent* of the following DE(s). The DE which follows the antecedent is an *anaphor*.

Table 1: Terminologies and examples

Term	Definition	Examples
DE	NP	*John*, *a book*, *it*, Ø
antecedent	DE which refer to the same concept or entity as other DE(s) and precedes other DE(s)	*John*, *a book*
anaphor	DE which follows its antecedent	*it*, Ø

In example (1), the zero pronoun indicated by Ø and the overt pronoun *it* are examples of anaphors, while *John* and *a book* are examples of antecedents of these anaphors. All of them are examples of DEs.

(1) **John**$_i$ bought **a book**$_j$ and
`antecedent1` `antecedent2`
Ø$_i$ read **it**$_j$.
`anaphor1` `anaphor2`

In this study, we do not include generic nouns such as *people in general* or generic *they* as in *they say that the war is over*. Our main interest is limited to anaphoric expressions and some discourse elements which refer to the speakers and the hearers in the narrative discourse. The terminologies are summarized in Table 1.

2.3. Issues specific to Japanese conversation

Several annotation schemes for English anaphoric relations have been proposed in the literature (Hirschman, 1997; Kingsbury and Palmer, 2002; Poesio et al., 2004; Doddington et al., 2004). The annotation scheme of coreference relations have been discussed in the Coreference (CO) tasks on Message Understanding Conference (MUC) and the Entity Detection and Tracking (EDT) task in the Automatic Content Extraction (ACE) program. As pointed out in Nakaiwa et al. (1995) and Iida et al. (2007), however, most of the (at least obviously) available cues for anaphors and antecedents in English are missing in Japanese because the most frequent type of anaphor in Japanese is zero pronoun. Thus, it is necessary to define a zero pronoun based on the argument structure before annotating anaphoric relations. As will be mentioned in the following section, we will employ *Thesaurus of Predicate-Argument Structure* (Takeuchi et al., 2010) to determine argument structure of the clauses and to identify zero pronouns. This methodology is suggested but has not been tried in Iida et al. (2010). In addition to the problems in zero pronominal languages discussed above, there are issues specific to conversations. Not only are there zero pronouns in Japanese, but also there are many kinds of zero pronouns especially in Japanese conversations. In (2), for example, there are at least six kinds of zero pronouns (Ø): (i) anaphoric zero pronouns indicated by the subscripts *i*, *j*, and *k*, (ii) those which refer to the speaker indicated by the subscript **sp**, (iii) those which refer to the hearer indicated by **hr**, (iv) those which refer to the participants as a whole in the conversation indicated by **pt**, (v) those which refer to something in the conversation setting (the room where they are talking, in this case) indicated by **ex** (which stands for exophora), and (vi) those which refer to something inferable from the context indicated by **inf**. Note that, except for *onee-san* 'lady' in B1, there are no explicit antecedents for these zero pronouns. This is extremely common in Japanese conversation. Also note that there is no obvious way to distinguish one kind from another; there is no verb agreement, and zero pronouns do not tell anything about their antecedents unlike *he* or *she* in English, which tells the gender and the number of the antecedent. For example, although *yuu* 'say' in B2 and A7 is identical in form, the subject is the speaker in B2 but the hearer in A7. It is perfectly fine to interpret the subjects in other ways if the context changes.

(2) B1: ano **onee-san**$_i$ kireeda-yone
that lady pretty-right
'That lady is pretty, isn't she?'
B2: ima **Ø**$_i$ **Ø**$_{\mathbf{ex}}$ detetta-kara **Ø**$_{\mathbf{sp}}$ **Ø**$_j$
now (she) (the room) left-because (I) (this)
yuu-kedo-sa
say-though-FP
'(I'm) saying (this) because (she) left (the room) now.'
ALL:⟨laugh⟩
C3: iya demo atti-de **Ø**$_i$ **Ø**$_{\mathbf{pt}}$ kiiteru-yo
no but over.there (she) (us) listen.to-FP
'No, wait. (She's) listening to (us) out there.'
ALL:⟨laugh⟩ ...
C4: a **Ø**$_k$ soo-ka
oh (that) right
'Oh, (that's) right.'
B5: **Ø**$_{\mathbf{hr}}$ kireena onee-san-wa sukidesu-ka
(you) pretty lady-TOP like-Q
'Do (you) like pretty ladies?'
C6: A, **Ø**$_{\mathbf{inf}}$ kawaii-to kiree-to dotti
A, (your favorite) cute-and pretty-and which
'Which is (your favorite), cute (ladies) or pretty (ladies)?'
A7: ee muzukasii koto **Ø**$_{\mathbf{hr}}$ yuu-ne
uh difficult thing (you) say-FP
'Uh... (You) say a difficult thing.'
`(chiba1032: 173.88-197.50)`

We expanded the methodology in Iida et al. (2007) as one of the earliest attempts to annotate anaphoric relations in Japanese and annotated anaphoric relations in Japanese conversations to find problems specific to spoken Japanese (and possibly other spoken languages which employ zero pronouns in a similar way). We do not claim that we could find the way to distinguish all of these types. Rather, we attempt to differentiate (i) anaphoric zero pronouns, (ii) the speaker, (iii) the hearer, and (v) exophora. The participants as a whole (iv) are difficult to tease apart from (ii) or (iii) as discussed in §5.7., and (vi) are also difficult to separate from some types of (i) and (v). We will mainly investigate the relationships between pronouns and DEs which occur explicitly in the conversations and leave the rest as open questions for further investigations.

3. Annotation scheme

We annotated 12 sessions of conversation from the *Chiba three-party conversation corpus* (Den and Enomoto, 2007). Each session is about 10 minutes, where three university students who know each other well talk about a topic which is triggered by rolling a dice with various topics on different sides. They can freely move on to another topic as the conversation goes on. The corpus has been annotated with various sorts of information such as morphological, prosodic, and clausal boundaries (Den et al., 2010).

The basic scheme for annotating anaphoric relations is as follows:

1. **Segmentation**: A sequence of discourse is segmented into a unit which corresponds to a clause. (§3.1.)
2. **Identification of argument structure and zero pronoun**: The argument structure of a clause is defined to decide what kind of zero pronoun occurs in the discourse. The arguments which are not overtly mentioned in the clause are determined to be zero pronouns. Zero pronouns in relative clauses or complement clauses are disregarded because they are argued to be different qualitatively from those in main clauses (Givón, 1983). (§3.2.)
3. **Identification and classification of DEs**: DEs in the discourse, including zero pronouns defined in the previous step, are identified and classified into categories based on what they refer to. (§3.3.)
4. **Identification of anaphoric relations**: For each DE identified in the previous step, the link between the anaphor and the antecedent is annotated if it refers to some antecedent. The candidates of antecedents are all DEs that occurred earlier in the discourse. (§3.4.)
5. **Annotation of topic continuity**: The information status and persistence of each DE are computed from the annotation of anaphoric relations in the previous step. (§3.5.)

3.1. Segmentation

Although we basically identify each clause as a unit of analysis, the following adjustment is made to better reflect the speakers' intuition. Although we decided to disregard zero pronouns in complement clauses, we regarded the complement clause of some psychological verbs such as *omou* 'think' and *ki-ga-suru* 'feel' as the main clause to take zero pronouns in those clauses into account. This is because the content of what the speaker thinks is more relevant to the flow of discourse rather than the event of the speaker's thinking itself.

3.2. Identification of argument structure and zero pronoun

We define zero pronouns based on the argument structure of a clause determined by *Thesaurus of Predicate-Argument Structure* (Takeuchi et al., 2010). The thesaurus employs Lexical Conceptual Structure (LCS) proposed in Jackendoff (1990). The argument of a clause does not include temporal, locative, and some kind of manner expressions. It provides alternative argument structure for ambiguous verbs, but does not decide which argument structure to apply. We manually modified arguments so that ambiguous cases are resolved.

We annotated surface case markers and did not annotate deeper argument relations such as thematic roles and underlying cases because surface case markers are expected to be more relevant to topic continuity (Givón, 2001). The valency changing operations such as passivizations and causativization are claimed to reflect topic continuity. Since the argument structure is defined in terms of basic verb forms in this thesaurus, valency altering operations are disregarded. If valency altering operations are observed in the corpus, we modified the argument structure to reflect the result of the operations. For example, the thesaurus identifies the subject (*ga*-NP) and the object (*o*-NP) for the verb *miru* 'see'. If the verb appears in the passive form *mir-areru* 'be seen', we erase the object and add *by*-NP (*ni*-NP). This is one of the differences between the current study and Iida et al. (2007). In their study deeper argument structures are annotated; they annotate anaphoric relations for automatic summarization and information extraction, where deeper argument structures are more useful.

In addition to zero pronouns, overt pronouns such as *kore* 'this' and *are* 'that', as well as *watasi* 'I', *omae* 'you' and variants of them are identified, which have antecedents somewhere else inside or outside the conversations.

3.3. Identification and classification of DEs

We distinguish four types of DEs: (a) the speaker and the hearer in the conversation ((ii) and (iii) in the last section for cases of zero pronoun), (b) the speaker and the hearer in a narrative discourse (subtypes of (ii) and (iii)), (c) post-predicative elements (special type of (i)), and (d) other anaphoric mentions, i.e., anaphors (typical case of (i)).

First, we identify the speaker and the hearer in the conversation. They are involved in communicative actions "here and now." Second, we identify the speaker and the hearer in narratives. They are involved in past or hypothetical events rather than communicative actions "here and now." Even though the same participants are involved in both of these, intuitively they are different and the participants in narratives are closer to topic pronouns. Third, we identify a DE which refers to the post-predicative element as in (3).

(3) **Ø**$_i$ tanosii-ne **ongaku**$_i$
(it) fun-FP music
'(It's) fun, I mean, music.'

`(chiba0332: 72.69-74.04)`

Although all kinds of NPs usually occur before the predicate in Japanese, post-predicate elements are frequently observed in spontaneous speech. We want to treat this in parallel with the case where place-holders such as *are* 'that' appear before the predicate as in (4).

(4) **are**$_i$ tukutteru-tte-yo ano (0.9) **PSX**$_i$
that make-QUOTE-FP FILLER (pause) PSX
'I heard that (he) is making that, I mean, PSX.'

`(chiba0232: 356.77-859.84)`

In both cases, we regard Ø and *are* 'that' as an antecedent of the post-predicative element.

Finally, we assume that other DEs that do not fall into any of the above categories are anaphoric; they are assumed to have antecedents somewhere in the discourse. The antecedents will be annotated in the following step.

3.4. Identification of anaphoric relations

Although it is important to distinguish *identify-of-reference anaphora* (IRA) and *identity-of-sense anaphora* (ISA) (Mitkov, 2002), we disregard this distinction so far. The distinction between IRA and ISA is exemplified in (5) and (6); the former is called IRA and the latter called ISA.

(5) **aru ten'in-san**-ga tikazuite-kite
some assistant-SUBJ approach-come
tiizu-baagaa-ni nari-masu-toka
cheese-burger-to become-POLITE-HEDGE
Ø itte gaan-te **Ø** oite
(she) say ONOMATOPOEIA-QUOTE (she) leave
Ø kaette-ki-masi-ta
(she) return-go-POLITE-PAST
'A shop assistant approached (to us), saying "here's cheeseburgers", left them to us, and went back.'
(chiba0332: 396.04-402.15)

(6) **medamayaki**-o tukutte-morau toki-ni itumo
fried.egg-OBJ cook-receive when-at always
ore-wa hanzyuku-ni site-tte itte-ndakedo
I-TOP sunny.side.up-to do-QUOTE say-though
itumo nazeka oya-ga koo
always somehow parent-SUBJ FILLER
kanzyuku-de **Ø** dasite-kuru-none
overcook-in (it) serve-come-FP
'Whenever I ask my mom to cook a fried egg, I ask her to make it sunny side up. But, somehow she always makes it overcooked.'
(chiba0232: 51.75-59.96)

In both (5) and (6), Ø can be interpreted as either the same entity that the antecedent refers to (IRA reading), or an entity of the same kind but not necessarily of the same entity (ISA reading). However, we know from the context that Ø refers to the same entity as the antecedent *ten'in-san* 'an assistant' refers to in (5), while it just refers to an entity of the same kind (not exactly the same entity) as *medamayaki* 'fried egg' in (6). Overt pronouns such as *sore* 'that' and *kore* 'this' are also ambiguous in the same way as zero pronouns.

We disregard this distinction because it is not reflected in linguistic forms.[1] We will annotate coreference relationships after coding anaphoric relationships.

3.5. Annotation of topic continuity

We propose topic continuity labeling for each DE using anaphoric relations identified in the preceding section. We distinguish two different annotations for topic continuity: *information status* and *persistence*.

Information status is relevant to the DE in question and its antecedent; in other words, the DE and the preceding discourse. According to Chafe (1994), there are three types of information statuses: *new*, *accessible*, and *given*. The information status of an entity is *new* if it has not activated in a person's mind when it is referred to in the discourse. It is *accessible* if the entity is activated in periphery of his or her mind. It is *given* if the entity is already activated. For our computational purpose, we define the information statuses of DEs as follows: the information status of a DE is *new* if there is no antecedent for the DE; *accessible* if there is another given or accessible DE intervening between the current DE and its antecedent; and *given* otherwise. Different DEs of different information statuses are expected to be realized as different linguistic forms (see Chafe (1994) for discussions mainly with respect to English).

As Givón (1983) argues, it is also important to see whether a DE is mentioned more than once in the following discourse; i.e., the relationship between the DE in question and the following discourse. We call this relationship *persistence* and distinguish two kinds of persistence: *persistent* and *non-persistent*. For our computational purpose, we define the persistence of DEs as follows: the DE in question is *persistent* if the current DE is an antecedent of some other DE, and it is *non-persistent* otherwise.

4. An illustrative analysis

4.1. Purpose

As a case study, we investigate the relationships between topic continuity and short-utterance units (SUUs) discussed in Den et al. (2010). An SUU constitutes a chunk of information that the speaker produces at a time. It is basically determined by acoustic and/or prosodic cues and corresponds to an intonation contour starting from the accentual phrase of bigger F_0-range to that of lower F_0-range. Figure 1 shows examples of SUUs. First, accentual phrases indicated by the solid black lines in the third row are determined. An accentual phrase begins with high (indicated by H in the second row), which gradually declines (indicated by L). Second, the F_0-range within an accentual phrase is calculated. Third, an SUU boundary is identified if the range of the following accentual phrase is bigger than the current accentual phrase. SUU boundaries are indicated by the solid gray lines in Figure 1.

Here we are especially interested in whether an overt DE and the predicate are uttered in separate SUUs or the same

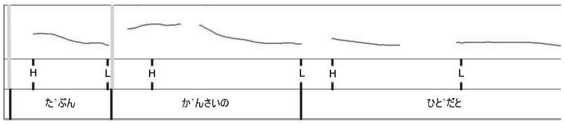

Figure 1: Example of SUU

[1] However, there is at least one expression *yatu* 'guys/things' which seems to be used exclusively for ISA reading (correspondence of English *one* as in *John got an iPad and Bill got one, too*), although the use of this expression is not obligatory. There are also other expressions such as *kooyuu* and *konna* 'this kind of', but they have to modify nouns. Most of the time, zero pronouns are used most frequently for both IRA and ISA.

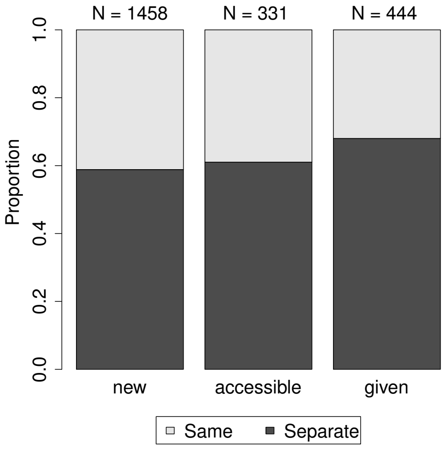

Figure 2: Information status vs. SUU

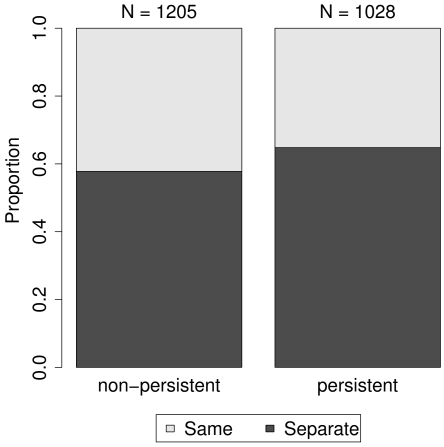

Figure 3: Persistence vs. SUU

SUU. Since an SUU corresponds to a chunk of information, features of topic continuity are expected to be related to SUUs. The literature such as Iwasaki (1993) points out that in Japanese conversation NPs and the predicate are separated prosodically more often than English, where usually a single intonation contour corresponds to a single clause (Chafe, 1994). We hypothesize that topic continuity features affect this difference.

4.2. Results

Figure 2 shows the distribution of DEs in terms of their information status and whether they appear in the same SUU as the predicate (*same*) or not (*separate*). It indicates that given DEs are separated from the predicate more often than others. Figure 3 shows the distribution of DEs in terms of their persistence and whether DEs appear in the *same* SUU with the predicate or in *separate* SUUs. It indicates that persistent DEs are separated from the predicate more often than non-persistent ones.

A logistic regression model was applied to study the interplay among information status, persistence, and the grammatical case which the DE occupies as an argument of the predicate, where speaker-based variance was considered as a random intercept. Model selection procedures based on AIC selected an optimal model with all the three fixed effects, the information status, the persistence, and the grammatical case. Table 2 shows the estimated coefficients of the fixed effects for the optimal model which predicts the same SUU realization. The possibility that *o*-marked (object) DEs appear in the same SUU as the predicate was significantly higher than *ga*-marked (subject) DEs ($p < .01$). The possibility that given DEs appear in the same SUU as the predicate was significantly lower than those of new DEs ($p < .05$); no difference was found between new and accessible DEs. Furthermore, the possibility that persistent DEs appear in the same SUU as the predicate was significantly lower than those of non-persistent DEs ($p < .05$).

These results suggest that topic-continuity features affect the production of NPs and the predicate with respect to prosodic structure, indicating efficacy of our annotation scheme in this kind of investigation.

Table 2: Logistic regression model predicting the same SUU realization (IS: Information status, P: persistence, GC: grammatical case)

	Coefficient	SE	z-value
Intercept	$-.39$	.08	-4.98***
IS=accessible	.01	.13	.06
IS=given	$-.29$	.12	-2.31*
P=persistent	$-.21$	.10	-2.21*
GC=o	.34	.11	3.06**
GC=ni	.19	.14	1.32
GC=to	$-.01$	.33	$-.02$
GC=wa	$-.39$	.25	-1.55

Reference levels: IS=new, P=non-persistent, GC=ga
*: $p < .05$, **: $p < .01$, ***: $p < .001$

5. Discussion

In this section, we discuss remaining issues in annotating anaphoric relations in spoken Japanese.

5.1. Additional layers

In addition to anaphoric relations annotated in this study, at least the following layers are expected to be useful for the study on topic continuity and discourse structure: (i) coreference relations, i.e., the relationships between DEs which refer to the same entity, not to the same concept, (ii) category-member relations, (iii) part-whole relations, and (iv) relations of identical linguistic forms.

Especially not considering (iv) was problematic in the annotation because sometimes the participants refer to two different entity which happen to have the same linguistic form in Japanese. In (7), for example, the topic triggered by the dice was about *koi* 'love affaires', but one of the participants started to talk about *koi* 'carp'. A1 triggers 'carp' by saying 'talking of *koi*', where *koi* refers to both 'love affaires' and 'carp'.

(7) A1: **koi**-no hanasi-tte ie-ba
carp/love.affairs-of topic-TOP say-if

are-da-ne
that-be-FP
'Talking of *koi*, by the way,'
A2: ano ibaraki-de **koi**-ga ippai sin-da-ne
um Ibaraki-in carp-SUBJ many die-PAST-FP
'Umm, a lot of carps died in Ibaraki.'
C3: ... a sotti-desu-ne
... oh that-POLITE-FP
'Oh, (you mean) that (*koi*).'
(chiba0732: 56.57-61.08)

Since we do not want to annotate anaphoric relations between 'love affaires' and 'carp', we simply exclude *koi* in A1 from the candidates of the antecedents of *koi* in A2.

5.2. One-word utterances

There are many one-word utterances in our corpus and it is not clear what kind of argument they take. Because spoken Japanese allows ellipsis of phrases as much as possible, it is often difficult to recover the full sentence. It seems necessary to distinguish at least the following kinds of one-word utterances: (i) copula predicates, where the subject NP is omitted, (ii) answers to questions, where NPs and VPs included in the question are omitted, (iii) corrections, additions, repetitions, or questions for the previous utterance, where unimportant NPs and VPs included in the previous utterance are omitted, and (iv) introductions to topic NPs without being associated with any full sentence.
An interesting type for our purpose is (iv). In (8), C1 mentioned *hakone* (a famous place for travel in Japan) out of blue. Since they had been talking about travel to Macedonia in the immediately preceding context, it is difficult to think of a missing subject. This might be a special way to introduce a topic and need to be described as such.

(8) C1: demo san-gatu **hakone**
but three-month Hakone
'By the way, March, Hakone!'
B2: **hakone**
Hakone
'Hakone!'
A3: o **sinkonryokoo**
oh honeymoon
'Honeymoon!'
B4: tigau
no
'No!'
ALL: ⟨laugh⟩
...
B5: sinkonryokoo-wa kaigai-tte
honeymoon-TOP oversea-QUOTE
kimeteru-mon
decide-FP
'(We) have already decided that (we're) going oversea for honeymoon.'
(chiba0632: 349.04-356.97)

5.3. Zero pronouns in complement clauses

Although we disregarded zero pronouns in complement clauses, sometimes complement clauses seem to be more relevant to topic continuity. It is necessary to identify zero pronouns in complement clauses as well to fully understand topic continuity. In (9), for example, the participants keep talking about *Satottyan* (a person's name), thus *Satottyan* is a continuous topic. However, zero pronouns, Ø$_i$, which refer to *Satottyan* in A2, B4, and C6 are inside the relative clauses and hence are disregarded based on our criteria.

(9) C1: **Satottyan**$_i$-wa nanka semer-areru-to
Satottyan-TOP FILLER be.active-PASSIVE-if
hontoni iya-sooda-yone
really dislike-appear-FP
'Satottyan appears to dislike other people to be active on him.'
...
A2: ano bimyoona **Ø**$_i$ tereru *kanzi*-ga
that subtle (he) blush atmosphere-SUBJ
kawaii-yone
cute-FP
'(His) blush-like atmosphere is cute.'
B3: Ø$_{\mathbf{sp}}$ **Ø**$_i$ wakan-nai-yone
(I) (him) understand-NEG-FP
'It's hard to understand (him).'
B4: tyotto uresi-sooni **Ø**$_i$ tereru *toki*$_j$-mo
a.bit happy-appear (he) blush time-also
aru-zyan
exist-FP
'Sometimes (he) blushes as if (he) is happy.'
C5: **Ø**$_j$ aru-ne
(that) exist-FP
'Year, (that) happens.'
...
C6: kekkoo aisode **Ø**$_i$ waratte kureru
often friendly (he) smile give
toki-ga aru
time-SUBJ exist
'Sometimes (he) smiles just to be friendly.'
...
C7: soko-o soo Ø$_{\mathbf{sp}}$ **Ø**$_i$ semete
that-OBJ that.way (I) (him) be.active
mita-kedo-ne
try-though-FP
'(I) tried to be active on (him), but ...'
(chiba1032: 527.90-559.80)

The claim that zero pronouns in complement clauses are qualitatively different from those in the main clause does not seem to apply to all Japanese complement clauses.

5.4. Disagreement with antecedents between participants

Sometimes the participants disagree what they refer to. In (10), A and B disagree with what they mean by *suiree* 'water-cooling one'. What A means by *suiree* is a water-cooling hard disk, while what B means by it is a cooler of the hard disk. The word *suiree* and the following zero pronoun refer to the same thing in the sense that A and B seem to pick up the antecedent from the previous discourse and successfully communicate with each other. On the other hand, they refer to different things in the sense that A and B means different things by *suiree*.

(10) B1: **suiree**$_i$ utteru-yo
water.cooling sell-FP
'Water cooling ones are sold.'

B2: **suiree**$_i$-no yunitto
water.cooling-of unit
'A unit of water cooling one.'

A3: Ø$_{\mathbf{sp}}$ **Ø**$_i$ sitteru
(I) (that) know
'(I) know (that).'

A4: **Ø**$_i$ NEC-no yatu-desyo
(that) NEC-of one-right
'(That's) the one made in NEC, right?'

B5: tigau tigau
no no
'No, no.'

B6: **sore**$_i$ hontai-desyoo
that main.part-right
'That's the main part.'

(chiba0232: 551.13-554.95)

Again, it is useful to have the linguistic expression layer in addition to the anaphoric relation layer as discussed in §5.1.

5.5. Unsuccessful anaphor

Sometimes the speaker is not successful to let the hearer pick up the right antecedent. The annotator will know the right antecedent by looking at the following discourse. However, this might not be a good annotation in the sense that the participants does not know the referent at the time an unsuccessful anaphor is uttered. For example, in (11), the anaphoric expressions *sore* 'that' in C3 and Ø$_i$ in C5 are unsuccessful judging from B's not answering to C's question. C keeps asking questions and finally, in C6, he says what he means by *sore* 'that' and Ø.

(11) B1: are HT-tekunorozii haitteru-to-sa
that HT-technology introduce-if-FP
'If HT technology is introduced into that,'

B2: benti-ga waruku-naru-yo
bench.mark-SUBJ worse-become-FP
'the bench mark gets worse.'

...

C3: e **sore**$_i$-tte doko-de kiiten-no
what that-TOP where activated-Q
'What? Where does it get activated?'

B4: un?
huh
'Huh?'

C5: **Ø**$_i$ dokode kiiteru-no
(it) where activated-Q
'Where does (it) get activated?'

C6: dakara **maruti-sureddo**$_i$-yo
so multi-threading-FP
'So, (I mean) multi-threading.'

(chiba0232: 427.55-447.68)

We need to define the failure of anaphora and annotate the failure as such.

5.6. Processing constraint

It is problematic to decide the antecedent when the anaphor is uttered too closely. We should not annotate cognitively unrealistic anaphoric relations. In Japanese conversations, the participants often repeat the same predicate to show their agreement. In (12), for example, the verb *aru* 'there is' is repeated many times. In our current scheme, we identify zero pronouns in the subject positions for all of the occurrence of the verb. This itself might be problematic. Moreover, in our current annotation, the antecedent of the first Ø$_i$ in C4 is annotated as Ø$_i$ in B3 and the antecedent of the first Ø$_i$ in A5 as the second Ø$_i$ in C4. Since they are speaking so fast and, thus, it is not realistic to think that C4 and A5 respond to the immediate preceding occurrence of Ø$_i$, it is necessary to set time constraints for antecedents to be valid.

(12) C1: ano zyaatte-sa koo aoi mizu-ga
that flushing-FP this.way blue water-SUBJ
nagareru **yatu**$_i$-desyo
flush thing-right
'That (bathroom) where blue water flushes, right?'

ALL: ⟨laugh⟩

A2: soo soo soo
yeah yeah yeah
'Year, exactly.'

B3: **Ø**$_i$ sinkan[sen-ka
C4: [**Ø**$_i$ aru [**Ø**$_i$ aru
A5: [**Ø**$_i$ aru **Ø**$_i$ aru

(chiba0532: 91.98-100.62)

5.7. Fusion of antecedents

As has been pointed out in Yamanashi (1992), sometimes an antecedent mentioned earlier is integrated into a larger category and is mentioned again. In (13), for example, one of the participants, C, is a freshman and does not do a part-time job. Another participant, A, is wondering whether other freshmen do not have part-time jobs. Then C enumerates all freshmen and whether they have part-time jobs or not.

(13) A1: **itinensee** minna baito-site-nai-n
freshman all part.time.job-do-NEG-Q
doona-no
how-Q
'Do all freshmen have no part-time jobs?'

C2: eetto **Naohiro**-wa sukunakutomo
FILLER Naohiro-TOP at.least
site-nai-desu-yo
do-NEG-POLITE-FP
'At least Naohiro doesn't have (one).'

...

C3: **hoka**-wa yattenzyanai-su-ka **minna**
other-TOP do-POLITE-Q all
'(I guess) all other (freshmen) have (one).'

A4: a de **Takuya** ima nani siten-no
oh then Takuya now what do-Q
'Oh, Takuya, what does (he) do now?'

(chiba1232: 339.35-349.98)

So far there is no way to represent this type of relationships. It is necessary to annotate category-member relations in addition to anaphoric relations as discussed in §5.1.

5.8. Participants in narrative

It is often ambiguous whether the speaker refers to the participants in narratives or "here and now." Similarly, it is not clear whether we need to identify anaphoric relations between the participants, or other DEs of the identical form, in two different narratives, told by the same speaker or different speakers. These problems will be partially solved by annotating coreference relations in addition to anaphoric relations.

6. Conclusion

This paper proposed a basic scheme for annotating anaphoric relations in Japanese conversations and discussed various kinds of problems specific to spoken languages mainly caused by on-line processing of discourse and/or interactions between the participants. This paper also proposed a method to compute topic continuity based on anaphoric relations. We showed that the topic continuity correlates with short-utterance units, which indicates the validity of our annotations of anaphoric relations and topic continuity and the usefulness for further studies on discourse and interaction.

For future studies, we will annotate more relationships between DEs as discussed in §5.1. and investigate a way to represent discourse structures.

7. Acknowledgements

This work is supported by Grant-in-Aid for Scientific Research (B) "Constructing cognitive and communicative models for dialog based on utterance-unit annotations" led by Yasuharu Den. We are grateful to the anonymous reviewers for their useful comments and suggestions.

8. References

W Chafe. 1994. *Discourse, Consciousness, and Time*. Chicago University Press, Chicago/London.

Y. Den and M. Enomoto. 2007. A scientific approach to conversational informatics: description, analysis, and modeling of human conversation. In T. Nishida, editor, *Conversational Informatics: An Engineering Approach*, pages 307–330. John Wiley & Sons, Hoboken, NJ.

Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: an empirically emerged scheme. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, pages 2103–2110, Valletta, Malta.

G. Doddington, M. Mitchell, L. Przybocki, S. Ramshaw, Strassel, and R. Weischedel. 2004. Automatic content extraction (ACE program): Task definitions and performance measures. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC2004)*, pages 837–840, Lisbon, Portugal.

T. Givón, editor. 1983. *Topic Continuity in Discourse*. John Benjamins, Amsterdam/Philadelphia.

T. Givón. 2001. *Syntax: An Introduction*, volume II. John Benjamins, Amsterdam.

K. Hashida, 2005. *GDA nihon-go anoteesyon manyuaru sookoo dai 0.74 ban (in Japanese)*. (GDA Japanese annotation manual: available at http://i-content.org/gda/tagman.html).

L. Hirschman, 1997. *MUC-7 coreference task definition. version 3.0*.

R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139, Prague, Czech Republic.

R. Iida, M. Komachi, N. Inoue, K. Inui, and Y. Matsumoto. 2010. Zyutugo-koo-koozoo-to syoooo-kankee-no anoteesyon: NAIST-tekisuto-koopasu-kootiku-no keekenkara (in Japanese). *Journal of Natural Language Processing*, 17(2):22–50. (Annotating predicate-argument relations and anaphoric relations: Findings from the building of the NAIST Text Corpus).

S. Iwasaki. 1993. The structure of the intonation unit in Japanese. *Japanese/Korean Linguistics*, 3:39–53.

R. Jackendoff. 1990. *Semantic Structures*. MIT Press, Massachusetts.

P. Kingsbury and M. Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the 3rd Language Resources and Evaluation Conference (LREC2002)*, pages 1989–1993, Las Palmas, Canary Islands, Spain.

D. McNeill, F. Quek, K. E. McCullough, S. Duncan, N. Furuyama, R. Bryll, X. F. Ma, and R. Ansari. 2001. Catchments, prosody and discourse. *Gesture*, 1(1):9–33.

R. Mitkov. 2002. *Anaphora Resolution*. Studies in Language and Linguistics. Longman, London.

H. Nakaiwa, S. Shirai, S. Ikehara, and T. Kawaoka. 1995. Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints. In *Proceedings of AAAI 1995 Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pages 99–105, Stanford, CA.

M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 144–151, Barcelona.

R. Sasano, D. Kawahara, and S. Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 769–776, Manchester.

K. Takeuchi, K. Inui, N. Takeuchi, and A. Fujita. 2010. A thesaurus of predicate-argument structure for Japanese verbs to deal with granularity of verb meanings. In *Proceedings of the 8th Workshop on Asian Language Resources*, pages 1–8, Beijing.

M. Yamanashi. 1992. *Suiron-to Syoooo*. Kuroshio, Tokyo. (Inference and anaphora).