TR-IT-0280

The Recognition of Noun Usage and Pronominal Anaphora in Japanese

- Towards Discourse Understanding Systems -

荒川直哉
ARAKAWA Naoya

November 1998

Abstract

This report is concerned with the referential aspects of nominals in Japanese. In particular, it deals with the automatic recognition of 1) noun usage and 2) pronominal anaphora. The classification of noun usage in the first part of this report includes the demonstrative, indexical, anaphoric, and so on. In the second part, the issue of automatic recognition of pronominal anaphora is addressed. Pronominal anaphora includes anaphora by pronouns and zero pronouns.

Table of Contents

Part 0. Introduction

This report is concerned with the referential aspects of nominals in Japanese. A referring expression in a discourse may introduce a new object or may denote a previously referred object or some object in the common knowledge of the discourse participants. Thus, a full-fledged natural language understanding system is required to tell these referential functions apart. For example, if a system is said to understand a discourse, it should be able to answer if an expression in the discourse denotes some object which has been denoted by another expression.

If more than one expression refers to an object in a discourse, the expressions are said to be *co-referential.* For an expression, the closest precedent co-referential expression (if any) is called the *antecedent,* and the two expressions are said to be *anaphoric* (the posterior expression is called *anaphor*). Finding anaphoric relations leads to the recognition of co-references necessary for understanding discourses.

This report consists of two parts. The first part deals with the recognition of noun usage and the second part the recognition of pronominal anaphora. In this report, nouns are classified according to how they refer to objects and what they refer to. The classification includes demonstrative, indexical, anaphoric, and so on. As mentioned above, it is important for a discourse understanding system to recognize these kinds of noun usage. In the second part, the issue of the automatic recognition of pronominal anaphora is addressed. Pronominal anaphora includes anaphora by pronouns and *zero pronouns*, which are non-realized (elliptical) obligatory deep case elements.

Part I. Statistical Recognition of Noun Usage in Spoken Japanese

This part discusses ways to recognise the usage of Japanese nouns, and describes experiments on a statistical recognition method. The part shows, through experiments, the effectiveness of the statistical recognition method and the importance of using background probabilities as well as lexical, syntactic and contextual clues. A travel domain dialogue corpus (the ATR-ITL Speech and Language DataBase: SLDB) was used for the experiments.

1. Introduction: Japanese Nouns

Japanese nouns often lack marks for definiteness even when their translations to another language are definite. This poses a problem in machine translation where the source language is Japanese and the definiteness in the target language should be explicitly marked. Moreover, the fact poses a difficulty for natural language understanding systems which need to determine whether nouns are anaphoric (i.e., if they denote previously mentioned objects).

Because of the above mentioned problems, a few classification methods have been proposed:

Table 1

Proposal	Classification
Murata et al. (1993)	genericity, definiteness and others; also by number
Bond et al. (1995)	genericity, ascriptivity, definiteness, countability, referential or not, and number
Siegel (1996)	definiteness and number
Heine (1998)	definiteness

All of these methods used heuristics for classification. The paper presented here classifies Japanese nouns into finer classes and uses statistics for classification. Moreover, while the previous works discuss the issue in relation to machine translation, the present work keeps the discussion within the mono-language framework, considering applications in mono-lingual dialogue systems.

2. Noun Usage

2.1 Specificity

Before discussing the usage of Japanese nouns, let us examine objects referred to by noun phrases. A discourse object, i.e., an object referred to in a discourse, can be specific or non-specific. A specific object is an identifiable object. In classical logic, a specific object would be represented with a logical constant. In languages like English, a specific object is referred to by a definite noun phrase in many cases, but also by an indefinite noun phrase when it is first introduced.

A non-specific object cannot be identified even by the speaker, and would be expressed as "something such and such" in English and as a logical variable in classical logic. A non-specific object is, in most cases, not subject to being an anaphoric antecedent, because it is, in terms of logic, supposed to exist as a variable only in the scope of a quantifier in a sentence. In some cases, it is not easy to judge whether an object referred to by a noun phrase is meant to be non-specific or introduced as a specific object. In the travel domain corpus which will be used in the experiment below[1], quite a few unspecific objects are referred to, because the objects of a reservation, such as rooms, are non-specific when the reservation task is in progress.

2.2 Generic noun phrases

The referent of a noun phrase may be a class of objects (e.g., the class of dogs). The subject of the English sentence "The dog barks." can refer to the class of dogs. Such reference is said to be *generic*. A generic expression referring to a specific class may be an indefinite noun phrase as in "Dogs bark.," as well as a definite noun phrase such as "the dog." The corpus used here does not have many generic examples, for it is a dialogue corpus in the travel domain where generic statements do not often appear.

2.3 Definite noun phrases and the modes of identification

In this paper, a noun phrase for which the speaker expects the hearer to identify its referent is said to be *definite*[2]. Thus, the referent of a definite expression is normally specific. Definite reference can be classified into various kinds according to the mode of identification, as explained below.

[1] "the corpus" henceforth

[2] cf. Arakawa (1995)

2.3.1 Anaphora

In many cases, the referent of a noun phrase can be identified by the fact that it has been introduced by a preceding expression in the discourse. Such reference is said to be an *anaphora.* The preceding expression A which denotes the referent of an expression B for the last time is called the *antecedent* of B. Here, B is said to be *anaphoric* to A.

2.3.2 Indexical

The referent of a noun phrase may be identified by the meaning of the noun phrase and the context where it has been uttered. For example, the referents of expressions such as "today," "here" and "I" can be identified by the meaning of the expressions and the situation in which they are uttered. Such use of expressions is said to be *indexical.*

2.3.3 Proper nouns

According to philosophers like Kripke (1980), proper nouns denote their referents by virtue of naming and the history of the name use. The referent of a proper noun is identified by the knowledge that it is called by the name (proper noun). Some (English) proper nouns, such as product names, may denote classes (e.g., the class of some commercial product).

2.3.4 Demonstrative

The speaker of a noun phrase may identify the referent by pointing to it. Such mode of identification is called *demonstrative* and often uses demonstratives such as "this" and "that."

2.3.5 Definite descriptions

The referent of a noun phrase may be identified solely based on the meaning of the noun phrase itself independent of the context of the utterance. For example, the referent of "the largest star in the universe" should be identified by the description independent of the context. In actuality, it is rare to find a definite description which can identify its referent without any contextual information; in most cases, we identify the referents of descriptions by both the meaning of the description *and* the information concerning the universe of discourse.

2.4 Predicative use

Traditionally, indefinite noun phrases such as that in "Hillary is a cat." have been analyzed as predicates. That is, the sentence above has been analyzed as a logical statement "cat(Hillary)." The predicative use is called *ascriptive* in Bond et al. (1995).

2.5 Others

There are other types of noun phrase usage which do not necessarily involve reference to specific objects.

Interrogatives

Interrogative noun phrases (interrogative pronouns and noun phrases modified by interrogative modifiers; e.g., "which book") do not denote specific objects.

Nouns which are actually modifiers

In Japanese, some words are classified as nouns but function only as modifiers used with the adnominal postposition "no."　(E.g., "*Betsubetsu*-no": separate)

Numeratives

In English, "two cups" as in "two cups of coffee" is a numerative.　They are a kind of quantifier and often denote non-specific objects.

Symbols

For example, alphabetic letters fall under this category.

Nominalized verbs

The infinitive and gerund of a verb can make up a noun phrase.

Idioms

Some noun phrases in idiomatic expressions do not have referents (e.g., "the sake" in "for *the sake* of" in English; "ため" in "のための" in Japanese).

Complementizers

Marks to indicate nominal clauses such as relative pronouns (e.g., "what") or the complementizer "that" in English.　In Japanese, "koto" is an example.

3. Clues to Recognize Noun Usage in Japanese

In this section, we are going to examine what kinds of clues can be used to recognize the usage of Japanese nouns.　Below, our consideration will be limited to common nouns (which do not include date/time expressions such as "Friday").　The clues to recognize common noun usage are lexical, syntactic or contextual, as explained below.

3.1 Lexical clues

The usage of some nouns can be judged lexically, i.e., without checking the context of their occurrences.　Judgment may be decisive or probabilistic.

3.1.2 Decisive clues

Nouns which are always complementizers, idiomatic or adjectival are lexically classified as such. For example, the nominal form of some verbs which are listed as nouns in the lexicon may always be treated as verbal. Letters are symbols. Words like “両方” (both) are inherently anaphoric.

3.1.2 Probabilistic clues

To classify nouns with many types of usage, the probabilities of their usage are used together with other clues (see 4.2). For example, the probability of the word “お部屋” (room)'s being indefinite is taken into account in the usage recognition process.

3.2 Syntactic clues

Some nouns may be judged to have a certain usage from their syntactic environment. For example, reference by a quantified noun is normally non-specific. Like lexical clues, syntactic clues can be decisive or probabilistic.

3.2.1 Decisive clues

Noun phrases modified by interrogative modifiers are interrogative. Common nouns following numbers are quantifying expressions (though they may be definite; e.g., “the last ten minutes”).

3.2.2 Probabilistic clues

Demonstrative modifiers are clues for the usage of the noun phrases they modify:

Table 3.1

Demonstrative modifier	Reference type
この (proximal)	demonstrative, indexical or anaphoric
その (mid-distal)	demonstrative or anaphoric
あの (distal)	demonstrative or reference to something in the “speaker's mind.”[3]

Expressions in an intensional context (a negative context or a volitional context) tend to be non-specific rather than specific. To draw an example in English, “a camera” in “Taro did not own a camera”[4] does not introduce a specific object to be referred to afterwards. Similarly, in phrases expressing non-actual situations such as “to buy a camera” in “Taro wants to buy a camera.,” noun phrases tend to be used non-specifically.

[3] In the corpus, the demonstrative use hardly occurs because it is a corpus of simulated *telephone* conversations.

[4] from Katagiri (1992)

Some other probabilistic clues are as follows. Modified noun phrases are more likely to be definite than otherwise. Expressions about the past would more likely be specific than non-specific, as descriptions of the past would tend to be factual. Expressions modified by quantifiers are more likely to be non-specific than specific. Here, quantifiers include numerative ones (indicating quantity) and logical ones such as "nanika-no" (some).[5]

3.3 Contextual clues

Contextual clues are used for finding anaphora: if a noun phrase has an antecedent candidate, it is likely to be anaphoric.

4. Experiments

Experiments were carried out on the automatic classification of common nouns. In the experiments, common nouns were classified using statistics from a corpus tagged with noun usage classification.

4.1 The Corpus

4.1.1 General information

The common nouns in the dialogue corpus were manually tagged with their usage. The corpus used was part of a travel domain corpus (SLDB) developed at ATR-ITL. The part contains 200 dialogues (7,650 utterances). The domain covers tasks such as making reservations for hotels, transportation, theaters and restaurants, as well as services at hotels.

4.1.2 Classification

Noun usage is classified as follows. The letters are used to mark each usage. (Most examples below are English equivalents.)

Definite nouns

X Non-personal indexical; e.g., "today"

D Demonstrative; e.g., "this room"

T Anaphoric; having antecedents

F Referent identified via background knowledge

f Referent identified via modifiers

d Other definite uses; such as reference to something in the speaker's mind or the contents of the preceding utterances (i.e., not anaphoric to specific expressions in the discourse)

[5] Logical quantifiers are not found in the corpus.

Indefinite nouns

U	Non-specific;	e.g., "I'd like to reserve a twin room."
P	Predicative;	e.g. "... is a twin room."
n	New object;	introducing new objects

Others

G	Generic	e.g., We are closed on Sundays.
Q	Interrogative;	e.g., "which book"
p	Adjectival;	e.g., "gogo" (pm)
q	Numerative;	e.g., two cups of coffee
s	Symbol;	e.g., "A" "B" "C"
V	Verbal;	verb derivatives
x	Idiomatic;	e.g., "for the sake of ..."
C	Complementizer;	noun clause markers; "koto," etc.

4.1.3 Statistics *(Table 4.1)*

X	116	2.23%
D	4	0.08%
f	727	13.98%
F	572	11.00%
T	554	10.65%
d	3	0.06%
DEF	1976	38.00%
U	1535	29.52%
P	217	4.17%
n	52	1.00%
INDEF	1804	34.69%
G	11	0.21%
Q	176	3.38%
p	415	7.98%
q	32	0.62%
s	2	0.04%
V	59	1.14%
x	53	1.02%
C	672	12.92%
Total	5200	100%

4.2 Algorithm

The algorithm is rather simple. First, the grammatical heads of noun phrases in a corpus are marked automatically with labels that are supposed to be relevant for noun usage recognition (4.2.1). Second, the probability of each noun usage relative to the marking patterns[6] is obtained from the training corpus (4.2.2). Third, in the classification stage, each common noun is given the label which has the greatest probability for its given marking pattern (4.2.3).

4.2.1 Marking

Lexical markers

- Semantic features

The following semantic features are given to each noun.

QUANT: quantifiers such as "すべて" (all) and "たくさん" (much)

LOC: location

HUM: human

SYMBOL: e.g., letters

SAHEN: *sahen* verbal nouns[7]

INDEXICAL: e.g., "きょう" (today), "あす" (tomorrow)

ANAPHORIC: always anaphoric;
e.g., "同日" (the same day)

- Semantic codes

The codes from a published thesaurus (Ôno et al. 1981) are given to the nouns in the corpus.

- Statistical properties

The following probabilities are given to each noun.

VERBAL: the probability of being V in the corpus

MODIFIER: the probability of being p in the corpus

DOMAIN_OBJE: the probability of being F in the corpus

NON-SPECIFIC: the probability of being U or P in the corpus

SPECIFIC: the probability of being F, f, T or n in the corpus

IDIOM: the probability of being x in the corpus

[6] E.g.,
```
p( X | QUANT+, HUM+, SPECIFIC=0.32, MODIFIED+)
    ↑ Usage   ↑ Marking pattern
```

[7] Nouns which transform into verbs by suffixing the auxiliary verb "suru"

COMPLEMENT: the probability of being C in the corpus

Syntactic markers[8]

WA: nouns marked by the topic marker "は" (wa). Since a topic tends to be a previously introduced object, a "wa"-marked noun would be more likely to be definite than otherwise.

ANAPHORIC: nouns modified by anaphoric modifiers such as "同" (the same) and "両" (both).

MODIFIED: nouns with modifiers

KONO: nouns modified by "この" (proximal demonstrative)

SONO: nouns modified by "その" (mid-distal demonstrative)

ANO: nouns modified by "あの" (distal demonstrative)

NO-MODIFIER: nouns modifying other noun phrases via the adnominal postposition "の"

PAST: nouns modified by predicates in the past tense; the case elements of predicates in the past tense

INTENSION: nouns in intensional contexts such as a negative predicate or a volitional context (e.g., case elements of the verbs with the volitional auxiliary verb "tai").

PREDICATIVE: Non-proper nouns before copula (The English sentence "A *is* a B" is translated into the Japanese sentence "A wa B *desu*" where "desu" is the copula.

WH-Q: nouns modified by interrogative

ANYTHING: nouns modified by indefinite modifiers such as "nanika-no" (any).

CLASSIFIER: Common nouns following numbers (like in "two *cups*").

QUANTIFIED: Existentially quantified nouns (e.g., "*heya*-ga aru/nai" [there is a/no room]); nouns with CLASSIFIER; nouns modified by counters (e.g. "huta-kago-no *ringo*" [two baskets of *apples*])

[8] The DEF/INDEF ratio varies with syntactic markers. The following table shows the DEF/INDEF ratio with some markers in the corpus.

Markers	Average = 1.26
WA	2.58
MODIFIED	1.64
PAST	2.29
INTENSION	0.93
QUANTIFIED	0.37

Antecedent candidates

The algorithm looks for antecedents for common nouns. Antecedent candidates are common nouns, date/time nouns, proper nouns, *sahen* nouns[9], and nominal compounds (ending with nouns or nominal-suffixes).

When an antecedent candidate is found, the corresponding common noun (anaphoric noun candidate) is marked with the type of the antecedent candidate. The types are as follows.

1. Exact string match: the antecedent has exactly the same form as the noun (marked as ANTE1).
2. Partial string match: the antecedent shares part of its form with that of the noun (marked as ANTE2). E.g. "ナンバー" (number) and "カードナンバー" (card number)
3. Semantic code matching: the antecedent has the same semantic code as the noun (marked as ANTE3).
4. Semantic feature matching: for generic place nouns ("ところ" "あたり" "場所"), a candidate with [LOC +] is sought. For the generic personal nouns ("本人" "連れ"), a candidate with [HUM +] is sought (marked as ANTE4).

The program first looks for an ANTE1 candidate in the preceding utterances.[10] If it fails, it looks for an ANTE2 candidate, and so on.

4.2.2 Statistics

For every marking pattern, the probability distribution for each noun usage is obtained. The occurrence of a *statistic property* with the probability **p** in a marking pattern is counted as its **p** times number of occurrence.

4.2.3 Classification

For each noun, the most probable type of noun usage is chosen based on its marking pattern and the statistics obtained above. Again, the occurrence of a *statistic property* with the probability **p** in a marking pattern is counted as its **p** times occurrence.

[9] In the experiment, nouns with [WH-Q +], [ANYTHING +], [PERSON 1], [PERSON 2], [VERBAL +], [SAHEN +] and [SYMBOL +] are excluded from the candidates of both anaphoric nouns and antecedents. Moreover, a noun whose probability of being marked as [IDIOM +] and [COMPLEMENT +] is higher than 0.8 is not considered as a candidate for an anaphor nor an antecedent.

[10] In the experiment, up to 40 utterances preceding anaphoric noun candidates (covering 442 [99.8%] anaphora in the corpus)

4.3 Results

The following two tables show the results for closed and open tests. `Ref.` stands for the number of usages given by human taggers. `Est.` stands for the number of calculated usages. `Match` stands for the number of matches between `Ref.` and `Est.`

Recall = `Match/Ref.`

Prec. (precision) = `Match/Est.`

`DEF` and `INDEF` stand for definite uses (`X`, `D`, `f`, `F`, `T`, d) and indefinite uses (`U`, `P`, `n`) respectively.

4.3.1 Closed test

The results are given in Table 4.2. Both test and training data reflect the entire tagged corpus. In Table 4.2, note that the `Match` numbers for `DEF` and `INDEF` are not the sums of the numbers for `Match` for the definite and indefinite uses respectively. They are the numbers of the cases where the definite/indefinite uses were correctly judged. This is also the case for the open test.

Table 4.2 (the closed test)

Class	Ref.	Est.	Match	Recall	Prec.
X	116	115	114	98.3%	99.1%
D	4	5	4	100%	80.0%
f	727	774	631	86.8%	81.5%
F	572	554	472	82.5%	85.2%
T	554	522	431	77.8%	82.6%
d	3	3	3	100%	100%
DEF	1976	1973	1788	90.5%	90.6%
U	1535	1647	1371	89.3%	83.2%
P	217	57	40	18.4%	70.2%
n	52	50	36	69.2%	72.0%
INDEF	1804	1754	1613	89.4%	92.0%
G	11	5	5	45.5%	100%
Q	176	174	174	98.9%	100%
p	415	417	402	96.9%	96.4%
q	32	33	32	100%	97.0%
s	2	2	2	100%	100%
V	59	59	59	100%	100%
x	53	48	48	90.6%	100%
C	672	735	663	98.7%	90.2%
Total	5200	5200	4486	86.3%	86.3%

4.3.2 Open test

The open test was carried out as a 5-way cross validation (i.e., the following result is the sum of five open tests where the corpus was divided into five parts and each part and the rest were used as the test and training data respectively). The results are given in Table 4.3. The explanation of `Match` for the closed test applies to the open test as well.

Table 4.3 (the open test)

Class	Ref.	Est.	Match	Recall	Prec.
X	116	95	94	81.0%	98.9%
D	4	2	0	0.0%	0.0%
f	727	757	576	79.2%	76.1%
F	572	568	454	79.4%	79.9%
T	554	468	342	61.7%	73.1%
d	3	1	0	0.0%	0.0%
DEF	1976	2034	1728	87.5%	85.0%
U	1535	1632	1305	85.0%	80.0%
P	217	69	34	15.7%	49.3%
n	52	43	29	55.8%	67.4%
INDEF	1804	1744	1543	85.5%	88.5%
G	11	6	4	36.4%	66.7%
Q	176	171	171	97.2%	100%
p	415	394	367	88.4%	93.2%
q	32	30	29	90.6%	96.7%
s	2	2	2	100%	100%
V	59	56	56	94.9%	100%
x	53	43	42	79.2%	97.7%
C	672	720	640	95.2%	88.9%
Total	5200	5200	4145	79.7%	79.7%

4.3.3 The Effects of Marking

Table 4.4 compares the prediction accuracy (=recall=precision) for experiments where groups of markers were used/not used. The background (bottom-line) accuracy (at the top) is shown for comparison. The accuracy rates are obtained by 5-way cross validation.

Table 4.4

	Accuracy
Background (bottom-line) (the probability of use U)	29.5%
Statistic Markers Not Used	51.1%
Non-Statistic Markers Not Used	66.3%
All Used	79.7%

The following tables show the effects of some markers. The right halves of the tables show the recall and precision for the most relevant types of usage. For example, Table 4.5 shows the results for the use T, when the contextual clues (ANTEs) were used and not used. The results were obtained by 5-way cross validation. The numbers in the top left-most cells stand for the number of occurrences of the markers in both training and test data.

Table 4.5

ANTE 2287	Total Accuracy	T Recall	T Precision
Used	79.7%	61.7%	73.1%
Not Used	79.0%	53.0%	61.5%

Table 4.6

MODIFIED 2244	Total Accuracy	f Recall	f Precision
Used	79.7%	79.2%	76.1%
Not Used	75.7%	53.0%	61.5%

Table 4.7

PREDICATIVE 667	Total Accuracy	P Recall	P Precision
Used	79.7%	15.7%	49.3%
Not Used	80.2%	9.2%	55.6%

Table 4.8

INTENSION 950	Total Accuracy	U Recall	U Precision
Used	79.7%	85.0%	80.0%
Not Used	81.1%	85.4%	79.4%

Table 4.9

QUANTIFIED	Total	U	U
481	Accuracy	Recall	Precision
Used	79.7%	85.0%	80.0%
Not Used	80.4%	85.1%	80.6%

Table 4.10

PAST	Total	DEF	DEF
45	Accuracy	Recall	Precision
Used	79.7%	87.5%	85.0%
Not Used	79.9%	87.5%	85.3%

Table 4.11

WA	Total	DEF	DEF
704	Accuracy	Recall	Precision
Used	79.7%	87.5%	85.0%
Not Used	80.2%	87.0%	85.6%

4.3.4 Antecedent resolution

Table 4.12 shows the recall rates of antecedent resolution in the cases where the algorithm could/could not recognize the anaphoric usage (T) correctly. The rates were calculated only for anaphora whose antecedents are single noun phrases.

Table 4.12

# of anaphora whose antecedents are single elements	443
# of correct antecedent choices (for correct usage estimation) Recall (235/443)	235 53%
# of correct antecedent choices (non-restrictive) Recall (317/443)	317 72%

4.4 Discussion

Table 4.4 shows that the usage statistics for each noun are important for usage recognition: lexical, syntactic, and syntactic clues alone do not yield good results.

In the tables showing the effects of lexical and syntactic markers, often the use of a marker brings about a minor loss in overall accuracy. This is due to the sparseness caused by the introduction of a new factor in the decision. That is, the more markers, the less data for each marking pattern with a given corpus.

Although the accuracy of the classification is lower than the results in Heine (1997), this is due to the complexity of the tasks with the corpus used (her corpus deals with only the scheduling task). Though heuristic rules in other works were examined, most of them were not decisive, and the weight for non-decisive heuristic rules should be statistically obtained for optimization as in this work.

Accuracy will be improved by improving the accuracy of antecedent resolution, although this will require an extensive use of knowledge.

References for Part I.

Arakawa N. (1995) *The Naturalization of Reference*, Temple University Dissertation.

Bond F., Ogura K. and Kawaoka T. (1995) *Noun phrase reference in Japanese-to-English machine translation*, In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI95), pp. 1-13.

Heine J. E. (1998) *Definiteness Prediction for Japanese Noun Phrases*, In Proceedings of COLING'98, pp. 519-525.

Katagiri Y. (1992) *状況意味論と談話理解 (Situation Semantics and Discourse Understanding)*, Journal of Japanese Society for Artificial Intelligence, 7/3, pp. 399-407 (in Japanese).

Kripke S. (1980) *Naming and Necessity,* Harvard University Press, Cambridge, Mass.

Murata M and Nagao M. (1993) *Determination of referential property and number of nouns in Japanese sentences for machine translation into English.* In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI93), pp. 218-225.

Ôno S. and Hamanishi M. (1981) Kadokawa Ruigo Shin-Jiten (角川類語新辞典), Kadokawa Shoten, Tokyo, 932p.

Siegel M. (1996) *Definiteness and Number in Japanese to German Machine Translation,* in "Natural language processing and speech technology : results of the 3rd KONVENS Conference, Bielefeld, October 1996", D. Gibbon, ed., Mouton de Gruyter, Berlin, pp. 137-142.

Part II. (Zero) Pronominal Anaphora Resolution

1. Introduction

1.1. The Issue

Pronominal anaphora resolution here means to find the antecedents (the closest preceding co-referential expressions) for (zero) pronouns among the precedent morphemes. Let us look at the following example in Japanese.

Example:

「今日のお昼にうな丼を学食で食べた。」 "I ate *unadon* for lunch at the school cafeteria today."
「それ、おいしかった？」 "Was *it* good?"

Morpheme sequence:
今日
の
お昼
に
うな丼 ← Antecedent
を
学食
で
食べ
た
それ ← Pronoun
…

In this example, the anaphora resolution for the pronoun「それ」is to select the proper antecedent「うな丼」in the morpheme sequence.

Various kinds of clues for anaphora resolution or antecedent selection have been proposed.[11] For example, one can expect that if the other conditions are the same, the closer the morpheme is to the pronoun, the more likely it is to be the antecedent. If an antecedent candidate is in the same sentence the pronoun is in, there are syntactic constraints on the candidacy. The Centering Theory (Grosz, Joshi, and Weinstein 1995) claims that grammatical functions such as Subject and Object have to do with the selection of the antecedent. Semantic affinities between the pronoun and morphemes should play a role too. For example, the antecedent of the pronoun「かれ」(he) would have the masculine gender and a singular number.

In this report, the case frame is considered to be an important key for resolving anaphora. Since a pronoun and the antecedent are co-referential, if the antecedent is a nominal, the replacement of the pronoun with the antecedent in the case frame (sentence) should be semantically fine. To take the example above, since「それ」is the subject in

[11] See the References for past research on pronominal anaphora resolution.

the case frame「おいしい」, the antecedent should also be able to be in the subject position for「おいしい」; i.e., the expression「うな丼はおいしい」should be semantically fine.

1.2. The Current Approach

A simple filtering program which judges if pairs of pronoun case frames and antecedent candidates would make up anaphora is used. For example, the program judges if the pair of (((それ（は <格助詞>)（おいし <形容詞>)) [the pronoun case frame] and (うな丼 <普通名詞>) [an antecedent candidate] is appropriate to be anaphoric. Automatic resolution is successful when the first antecedent candidate which is judged to be the antecedent is actually the antecedent. The evaluation of the candidates starts from the utterance previous to the utterance where the pronoun occurs and recedes along the time line (in the experiments, only inter-sentential anaphora resolution was tested). In each utterance, kinds of intra-sentential ordering of the candidates are tested (see 3.3).

For the judgment of the appropriateness of the pronoun-antecedent candidate pairs, a statistical method was used: the probabilities of the pronoun-antecedent candidate pairs and the replacement of antecedent candidates into the pronoun case frames were calculated from the statistics in a corpus.

2. The Scope of the Experiments

In this section, the 1) anaphora data for statistics and evaluation, 2) qualifications on the experiments, and 3) auxiliary data employed will be explained.[12]

2.1. Anaphora Corpora

The main data for the experiments is a corpus of travel conversations tagged with the usage and anaphoric relations on pronouns and zero pronouns. 375 dialogues in the ATR-ITL Speech and Language DataBase (SLDB) tagged for this purpose were used.

2.2. Qualifications

Not all of the anaphoric data in the corpus was used. The following is the list of qualifications with regard to the data:

- Only anaphoric pronouns are used: those for demonstrative or indexical use, for example, are excluded.[13]
- Only inter-sentential anaphora are examined: intra-sentential anaphora are excluded.
- Only those with single antecedents are used. Moreover, the pronouns "どちら," "どっち," "いずれ," "どれ," and "それぞれ" are excluded, as they are assumed to have

[12] See the Appendix for the location of the data.

[13] A real system that resolves anaphora should distinguish pronouns for anaphora use from those for other uses. For the automatic classification of pronoun uses, refer to (Yamamoto et al. 1998).

multiple antecedents.

- Only those with nominal antecedents are used: those with predicative antecedents, for example, are excluded. The nominal in this report is defined by the following parts of speech:

  (<普通名詞> <固有名詞> <代名詞> <名詞句> <接尾辞> <サ変名詞>
  <数詞> <日時> <副助詞> <人名> <住所名> <副詞的名詞>)

2.3. Auxiliary Data:

- **Case Analysis Corpus**

As mentioned above, case frames play important roles in the experiments. The case frames were taken from the case analysis corpus created at ATR-ITL for the SLDB dialogues.

- **Dictionaries for Semantic Information**

To make anaphora resolution robust, semantic abstraction by semantic codes (semcodes, hereafter) and semantic features (HUM, LOC, TLOC) is used.

Semcodes:

A semcode list has been prepared at Department 3 of ATR-ITL following the system of The Kadokawa Thesaurus (大野 et al. 1981). <日時> (temporal) expressions not included in the list are added manually.

Semantic Features:

A semantic feature dictionary was build manually.

3. Details

Although the basic idea for the experiments is the same for pronouns and zero pronouns, there are minor differences. Below, details unique to regular pronouns are explained first, and then those unique to zero pronouns are explained. Finally, details common to both regular pronouns and zero pronouns are explained.

3.1. Pronoun Anaphora Resolution

3.1.1. The pronoun-antecedent candidate pair

As mentioned earlier, the filter program judges if pairs of pronoun case frames and antecedent candidates are appropriate to be anaphoric. The following is an example pair and the format of pronoun-antecedent candidate pairs. Although the antecedent candidate is associated with a case frame, its predicative (verb or adjective) was not actually used.

A sample pair:

```
TCC22071-0150-1,15,TCC22071-0140-1,11,
(((それ (を <格助詞>) (作れ <本動詞> 一段))
 ((ごはん <普通名詞>) (を <格助 詞>) (炊 <本動詞> 五段カ)) OBJE) -),C
```

The format:

```
1. the utterance ID of the pronoun
2. the sub-utterance index of the pronoun
3. the utterance ID of the candidate
4. the sub-utterance index of the candidate
5. (((pronoun
     second-case-frame-component [in most cases a 助詞 (case marker)]
     third-case-frame-component  [in most cases a verb])[14]
    (first-case-frame-component
     second-case-frame-component
     third-case-frame-component)
   CASE)      ; the thematic role (deep case) of the pronoun
  polarity)   ; "+" if antecedent, else "-"
6. C iff the COND case in a TFQ or a confirmation question [optional]
```

3.1.2. Screening Process

The pronoun-antecedent candidate pairs which are judged inappropriate to be anaphoric are screened out. Statistic screening and hand-coded constraints were used in the experiments. There are two kinds of statistic screening: one uses positive and negative examples of pronoun-antecedent candidate pairs, and the other uses the case frame database. Moreover, hand-coded heuristics that pass certain kinds of pronoun-antecedent candidate pairs as appropriate were used.

3.1.2.1. Screening by Learning

Pronoun-antecedent candidate pairs are used to learn which pairs are likely to be anaphoric. The anaphoric pairs are positive examples and those without actual antecedents are negative examples.

Pronoun-Antecedent Combination Check

If the combination of a pronoun and a candidate nominal often appears in the set of pronoun-antecedent candidate pairs but mostly as a negative example, then the combination is screened out. The candidate is matched by the morpheme (the form & part of speech) or only by its part of speech.

Example pattern:	(それ (ごはん <普通名詞>))	morpheme matching
	(それ <普通名詞>)	part-of-speech matching

[14] When the pronoun occurs in the construction PRON-の-NOMINAL, the case frame contains (PRON (の <連体助詞>) NOMINAL). When the pronoun occurs in a copular construction such as PRON-は-NOMINAL-だ, the case frame is represented as (PRON (は <係助詞>) NOMINAL).

Case Frame Check

If the substitution of the antecedent for the pronoun in a case frame often appears in the set of pronoun-antecedent candidate pairs but mostly as a negative example, then the combination is screened out.

Example pattern: ((それ (を <格助詞>) (作れ <本動詞> 一段)) (ごはん <普通名詞>))

Substitution –>((ごはん <普通名詞>) (を <格助詞>) (作れ <本動詞> 一段))

Patterns are matched via semcode as well. E.g.,

(("354" "922") (を <格助詞>) ("260" "391" "396" "808a"))

Case frames using only 助詞 without the predicative are also used. E.g.,

((ごはん <普通名詞>) (を <格助詞>))

(("354" "922") (を <格助詞>))

3.1.2.2. Screening with Case Frame Database

If the substitution of the antecedent for the pronoun in a case frame does not appear enough in the case frame database compared to the occurrence of the case frame without the pronoun part, the pair is screened out. Because sparseness is expected, semantic features and parts of speech are used instead of morphemes.

Example: If the case frame (((TLOC +)) (へ <格助詞>) (着 <本動詞> 五段カ)) does not appear enough in the case frame database compared to the frame without the pronoun part ((へ <格助詞>) (着 <本動詞> 五段カ)), it is screened out.

3.1.2.3. Constraints

The following non-statistic constraints were used to judge certain kinds of pronoun-antecedent candidate pairs as inappropriate to be anaphoric.

Nouns in Modifier Use Check

Nouns used as modifiers such as "特別," "別々," "本当" are screened out from the antecedent candidates with a hand-made dictionary.

Pronoun-Antecedent Feature-Conflict Check

Feature conflicts between pronouns and antecedent candidates are handled by hard coding (↓). (These constraints can be learned and the hand-coding will turn out to be unnecessary [see 5. Results]. The constraints are set weak so that they do not screen valid combinations out.)

```
(case (first (first pattern))
      ((あそこ ここ そこ) '((HUM +) (LOC -) (TLOC +)))
      ((あちら あっち) '((TLOC +) (HUM +)))
      ((おたく) '((TLOC +)))
      ((それ) '((HUM +)))
      ((これ どれ) '((HUM +) (TLOC +)))
      ((それぞれ) NIL)
      ((こちら どちら) NIL)
      ((そちら) '((HUM +) (TLOC +)))
      ((こっち そっち どっち) '((HUM +)))
      ((彼) '((HUM -) (LOC +) (TLOC +)))
      ))
```

LISP code describing conflicts between pronouns and features

3.1.2.4. Heuristics

The following heuristics were used to pass certain kinds of pronoun-antecedent candidate pairs as appropriate to be anaphoric. (The heuristics turned out to be ineffective in the experiments.)

Same Pronoun Check

If the antecedent candidate is a pronoun of the same form, it is judged to be the antecedent.

COND in the Previous TFQ/Confirmation Check

If 1) the pronoun has the thematic role COND (CONDition) and 2) the antecedent candidate is in the previous utterance and of the type TFQ (True-False Question) or Confirmation-Question and has the thematic role COND, then the candidate is judged to be the antecedent.

3.2. Zero Pronoun Anaphora Resolution

3.2.1. The pronoun-antecedent candidate pair

As mentioned earlier, the filter program judges if pairs of (zero) pronoun case frames and antecedent candidates are appropriate to be anaphoric. The following is an example pair and the format of pronoun-antecedent candidate pairs. Although the antecedent candidate is associated with a case frame, its predicative (verb or adjective) was not actually used.

A sample pair:

```
TAC22013-0090-1,16,TAC22013-0020-1,5,
((((願 <本動詞> 五段ワ) OBJE)
  ((予約 <サ変名詞>) (を <格助詞>) (願 <本動詞> 五段 ワ))) +)
```

The format:

```
1. the utterance ID of the pronoun
2. the sub-utterance index of the pronoun
3. the utterance ID of the candidate
4. the sub-utterance index of the candidate
5. (((predicative thematic-role)
    (first-case-frame-component
     second-case-frame-component
     third-case-frame-component))
   polarity) ; "+" if antecedent, else "-"
```

3.2.2. Screening

3.2.2.1. Screening by Learning

Pairs of zero pronouns with their case frames and their antecedent candidates can be used to learn which pairs are likely to be zero pronoun-antecedent pairs. The pairs with actual antecedents are positive examples and those without actual antecedents are negative examples.

Zero Pronoun-Antecedent Combination Check

If the combination of the thematic role of the zero pronoun and the candidate often appears in the corpus of zero pronoun-antecedent candidate pairs but mostly as a negative example, then the combination is screened out. The candidate is matched with its form, semcodes, semantic features, or parts of speech.

Example patterns:

```
(DEST (金曜 <日時>))
(DEST ("151d"))
(DEST ((TLOC -)))
(DEST <日時>)
```

Case Frame Check

If the embedding of the antecedent to the zero position of the case frame often appears in the corpus of zero pronoun-antecedent candidate pairs, but mostly as a negative example, then the combination is screened out. A pattern matched with morphemes, semcodes, semantic features, or parts of speech.

Example patterns:

```
((金曜 <日時>) DEST (着 <本動詞> 五段カ))      Morpheme Matching
(("151d") DEST ("223" "302b" "314b" "386c")) Semcode Matching
(((TLOC -)) DEST (着 <本動詞> 五段カ))        Sem.Feature Matching
(<日時> DEST (着 <本動詞> 五段カ))            Part-of-Speech Matching
```

3.2.2.2. Screening with Case Frame Database

A pair of a zero-pronoun case frame and an antecedent candidate is screened out if the embedding of the antecedent to the zero does not appear enough in the case frame database compared to the occurrence of the case frame itself. Because sparseness is expected, semantic features and parts of speech are used instead of morphemes. The case frame database is constructed from zero-pronoun antecedent pairs.

Example: If the case frame (((TLOC +)) DEST (着 <本動詞> 五段カ)) does not appear enough in the case frame database compared to the frame (DEST (着 <本動詞> 五段カ)), it is screened out. Similarly, if the case frame (<日時> DEST (着 <本動詞> 五段カ)) does not appear enough in the case frame database compared to the frame (DEST (着 <本動詞> 五段カ)), it is screened out.

3.2.2.2. Constraints

Nouns in Modifier Use Check

As with regular pronouns, nouns that are used as modifiers such as "特別," "別々," "本当" are screened out from the antecedent candidates with a hand-made dictionary.

Zero Pronoun-Antecedent Feature Conflict Check

Feature conflicts between zero pronouns (case slots) and antecedent candidates were handled by hard coding. (These constraints can be learned, but full learning requires more data. The constraints are set weak so that they do not screen valid combinations out.) (Ref. ↓)

```
(case (second (first pattern))
      ((AGEN) '((TLOC +)))
      ((CONT) '((TLOC +) (HUM +) (LOC +)))
      ((DEPT DEST LOCT) '((HUM +) (TLOC +)))
      ((EXPR) '((TLOC +)))
      ((MUTL) '((TLOC +)))
      ((GOAL) '((HUM  +)))
      ((RECP) '((TLOC +)))
      ((ROUT) '((HUM +) (TLOC +)))
        ))
```

LISP code describing conflicts between zero pronouns and features

3.3. Intra-sentential Ordering

Three kinds of intra-sentential ordering were tested. In the standard setting, if one condition did not bring about ordering, then the next condition was used for ordering.

3.3.1. Ordering by Semantic Affinity

In this ordering, priority is given to a pair where the substitution of the antecedent candidate for the pronoun (for a zero pronoun, the embedding of the antecedent candidate to the zero position of the frame) occurs more in the case frame database.

3.3.2. Ordering by Grammatical Function

In this ordering, priority is given based on the Centering Theory[15]. Namely, 係助詞 (TOPIC) > を > に > 格助詞 (Other Case Markers) > Others. Zero pronouns (see 3.4) get a higher priority than other elements (i.e., AGEN > OBJE > ZERO > 係助詞) under the assumption that zero pronouns are used to refer to the most salient objects.

3.3.3. Ordering by the Order of Occurrence

In this ordering, priority is given to an element that occurs later (i.e., closer to the pronoun). This is the default ordering.

3.4. Data Reliability & Thresholds

The screening methods mentioned above use the numbers of occurrences of syntactic patterns. When the corpus used is not large enough, these numbers may not represent the probabilities of occurrences. Roughly speaking, when a probability is calculated as a fraction, the larger the denominator is, the higher the reliability of the calculated probability can be. Thus, the experiments were set so that screening would be done only when the denominators are larger than certain thresholds (which depended on the types of screening). The thresholds were empirically determined.

4. Results

4.1. Pronoun Anaphora Resolution

Below, A-E each signifies a different way for 5-way cross validation tests. The scores are success rates from anaphora resolution. The "Total" scores on the right are cross validation summations. The nine sets of scores following "The Best Score" are obtained by removing one screening process or heuristic or adding one kind of intra-sentential ordering from "The Best Score" anaphora resolution process. The last set of scores shows the baseline, namely, the score obtained when the program judges the closest precedent nominals to be the antecedents of pronouns. By default, intra-sentential ordering is done only by the occurrence order (ordering by semantic affinity & grammatical functions is not used).

- The Best Score

A	B	C	D	E	Total
.5821 (39/67)	.6552 (38/58)	.6491 (37/57)	.6301 (46/73)	.5733 (43/75)	.6152 (203/330)

- Pronoun-Antecedent Combination Check (Case-Based) Off

A	B	C	D	E	Total
.5224 (35/67)	.5345 (31/58)	.5439 (31/57)	.6164 (45/73)	0.52 (39/75)	.5485 (181/330)

[15] The Centering Theory by Grosz, Joshi and Weinstein (1995) concerns the preference by grammatical functions only in the previous utterance, while the ordering here is not limited to the directly preceding utterance.

- Case-Frame Check (Case-Based) Off

A	B	C	D	E	Total
.5373 (36/67)	.6379 (37/58)	0.614 (35/57)	.6164 (45/73)	.5867 (44/75)	.5970 (197/330)

- Screening with Case Frame Database Off

A	B	C	D	E	Total
.5821 (39/67)	.6724 (39/58)	.6491 (37/57)	.6301 (46/73)	0.56 (42/75)	.6121 (202/330)

- Nouns in Modifier Use Check Off

A	B	C	D	E	Total
.5672 (38/67)	.6552 (38/58)	.6491 (37/57)	.6301 (46/73)	.5733 (43/75)	.6121 (202/330)

- Pronoun-Antecedent Feature Conflict Check Off

A	B	C	D	E	Total
.5821 (39/67)	.6724 (39/58)	.5965 (34/57)	.6027 (44/73)	.5733 (43/75)	.6030 (199/330)

- Same Pronoun Check Off

A	B	C	D	E	Total
.5821 (39/67)	.6552 (38/58)	.6491 (37/57)	.6301 (46/73)	.5733 (43/75)	.6152 (203/330)

- COND in the Previous TFQ/Confirmation Check Off

A	B	C	D	E	Total
.5522 (37/67)	.6552 (38/58)	.6491 (37/57)	.6301 (46/73)	.5733 (43/75)	.6091 (201/330)

- Intra-Sentential Ordering by Semantic Affinity On

A	B	C	D	E	Total
.5672 (38/67)	0.5 (29/58)	.6667 (38/57)	.6164 (45/73)	0.6 (45/75)	.5909 (195/330)

- Intra-Sentential Ordering by Grammatical Function On

A	B	C	D	E	Total
.5821 (39/67)	.6207 (36/58)	.6842 (39/57)	.5479 (40/73)	0.52 (39/75)	.5848 (193/330)

- Non-Stochastic Checks [Nouns in Adj., Feat. Conflict, Same Pron. & COND-TFQ] Off

A	B	C	D	E	Total
.5373 (36/67)	.6724 (39/58)	.5965 (34/57)	.6027 (44/73)	.5733 (43/75)	.5939 (196/330)

- Stochastic Checks Off

A	B	C	D	E	Total
.4776 (32/67)	.5517 (32/58)	.4912 (28/57)	.5890 (43/73)	.4933 (37/75)	.5212 (172/330)

- Baseline: All Checks Off (Linear Order Only)

A	B	C	D	E	Total
.4328 (29/67)	.5517 (32/58)	.4035 (23/57)	.5342 (39/73)	.4933 (37/75)	.4848 (160/330)

- Closed Test

The success rate for the closed test with the same conditions as the cross validation (open experiment) with the best score is 0.7273 (240/330).

- Summary

Due to redundancies in screening, removing one kind of screening does not bring about much deterioration of the result. The Case-Frame Check (by a Case-Frame Database) and the Same Pronoun Check can be safely removed. The efficacy of the Nouns in Modifier Use Check, the Pronoun-Antecedent Feature Conflict Check, and the COND in the Previous TFQ/Confirmation Check seems small (if any).

Intra-sentential reordering by semantic affinity or grammatical functions has negative effects in the experiments above.

- Success Rates with regard to Pronouns (the best cross validation score)

Pron.	Success Rate
あそこ	0.0 (0/1)
あちら	1.0 (1/1)
ここ	.857 (6/7)
こちら	0.5 (51/102)
こっち	0.0 (0/1)
これ	.652 (15/24)
そこ	0.64 (32/50)
そちら	.606 (20/33)
そっち	0.0 (0/1)
それ	.704 (76/108)
彼	0.5 (1/2)

4.2. Zero Pronoun Anaphora Resolution

A-E each signifies a different way for 5-way cross validation tests. The scores are success rates from anaphora resolution. The "Total" scores on the right are cross validation summations. The seven sets of scores following "The Best Score" are obtained by removing one screening process, heuristic, or one kind of intra-sentential ordering from "The Best Score" anaphora resolution process. The last set of scores shows the baseline, namely, the score obtained when the program judges the closest precedent nominals to be the antecedents of zero pronouns.

- The Best Score

A	B	C	D	E	Total
.5354 (257/480)	.5131 (216/421)	.4322 (236/546)	.4953 (261/527)	.4959 (241/486)	.4923 (1211/2460)

- Zero Pronoun-Antecedent Combination Check (Case-Based) Off

A	B	C	D	E	Total
.4792 (230/480)	.4988 (210/421)	.3773 (206/546)	.4611 (243/527)	.4753 (231/486)	.4553 (1120/2460)

- Case-Frame Check (Case-Based) Off

A	B	C	D	E	Total
.4813 (231/480)	.4608 (194/421)	.3938 (215/546)	.4497 (237/527)	.4383 (213/486)	.4431 (1090/2460)

- Screening with Case Frame Database Off

A	B	C	D	E	Total
.5333 (256/480)	.5036 (212/421)	.4121 (225/546)	.4820 (254/527)	.4753 (231/486)	.4789 (1178/2460)

- Nouns in Modifier Use Check Off

A	B	C	D	E	Total
.5333 (256/480)	.5107 (215/421)	.4322 (236/546)	.4934 (260/527)	.4959 (241/486)	.4911 (1208/2460)

- Zero Pronoun-Antecedent Feature Conflict Check Off

A	B	C	D	E	Total
.5333 (256/480)	.5131 (216/421)	.4322 (236/546)	.4934 (260/527)	.4959 (241/486)	.4915 (1209/2460)

- Intra-Sentential Ordering by Semantic Affinity Off

A	B	C	D	E	Total
.5188 (249/480)	.4964 (209/421)	.4267 (233/546)	.4877 (257/527)	.4856 (236/486)	.4813 (1184/2460)

- Intra-Sentential Ordering by Grammatical Function Off

A	B	C	D	E	Total
.5375 (258/480)	.5202 (219/421)	.4231 (231/546)	.4972 (262/527)	.4877 (237/486)	.4907 (1207/2460)

- Non-Stochastic Checks [Nouns in Adj., Feat. Conflict] Off

A	B	C	D	E	Total
.5312 (255/480)	.5107 (215/421)	.4322 (236/546)	.4915 (259/527)	.4959 (241/486)	.4902 (1206/2460)

- Stochastic Checks (+ Intra-Sentential Ordering by Semantic Affinity) Off

A	B	C	D	E	Total
.2708 (130/480)	.3017 (127/421)	.2161 (118/546)	.2638 (139/527)	.2778 (135/486)	.2638 (649/2460)

- Baseline: All Checks Off (Linear Order Only)

A	B	C	D	E	Total
.1750 (84/480)	.2399 (101/421)	.1465 (80/546)	.1841 (97/527)	.1872 (91/486)	.1841 (453/2460)

- Closed Test

The success rate for the closed test with the same conditions as the cross validation (open experiment) with the best score is 0.7663 (1885/2460).

- Summary

Due to redundancies in screening, removing one kind of screening does not bring about much deterioration of the result. Unlike in the pronoun experiments, intra-sentential-ordering by semantic affinity or grammatical functions does not have negative effects.

5. Discussion

With more data and heuristics, one may expect the result to be near 70% (see the results for the closed tests). However, close examination of the result (See the Appendix) seems to tell us that "deep knowledge" is required to attain more. As for heuristics, it may be of help to detect parallelism in the preceding utterances.

References for Part II

Abraços et al. (1994) "Extending DRT with a Focusing Mechanism for Pronominal Anaphora and Ellipsis Resolution," In Proceedings of COLING'94, pp. 1128-1132.

荒川直哉 (1995) "代名詞の先行詞推定に関するセンタリング理論の評価," ATR-ITL Technical Report TR-IT-0141.

Grosz B., Joshi A. and Weinstein S. (1995) "Centering: A Framework for Modeling the Coherence of Discourse," *Computational Linguistics*, Vol. 21, #2, pp. 203-225.

Lappin S. and Leass H. (1994) An Algorithm for Pronominal Anaphora Resolution, Computational Linguistics, Vol. 20, #4, pp. 535-561.

村田真樹、長尾真 (1998) "表層表現と用例を用いた照応省略解析手法," 信学技報 (Technical Report of IECE), NLC97-56, pp. 9-16.

中岩浩巳 (1998) "日英対訳コーパス中のゼロ代名詞とその指示対象の自動認定," 自然言語処理 123-5, pp. 33-40.

Nasukawa T. (1994) "Robust Method of Pronoun Resolution Using Full-Text Information," In Proceedings of COLING'94, pp. 1157-1163.

大野晋、浜西正人 (1981) *角川類語新辞典*, 角川書店, Tokyo, 932pp.

Takada S. and Doi N. (1994) "Centering in Japanese: A Step Towards Better Interpretation of Pronouns and Zero-Pronouns," In Proceedings of COLING'94, pp. 1151-1156.

田村浩二、奥村学 (1995) "センター理論による日本語談話の省略解析," 自然言語処理 107-12, pp. 91-96.

Walker M., Iida M. and Cote S. (1994) "Japanese Discourse and the Process of Centering," *Computational Linguistics*, Vol. 20 #2, pp. 193-231.

Yamamoto K. and Sumita E. (1998) "Feasibility Study for Ellipsis Resolution in Dialogues by Machine-Learning Technique," In Proceedings of COLING'98, pp. 1428-1477.

Part III. Appendix

A1 Summary of Tags

(PRON: SHOUOU.DAT.NEW
ZERO: *.zero-conv
NOUN: *.hutu-conv)

A : (PRON) Reference to propositions in multiple utterances
Fields 3 & 4 specify the closest elements of the antecedent, but there are also unspecified elements.

A : (ZERO) Antecedents are distributed widely in the discourse.
The closest antecedent is specified after "A;".

B : (PRON) The anaphoric element (anaphor) is part of a conjunctive (e.g., "それに").

C : (PRON) The antecedent is part of a complex noun
「ニューヨーク観光局」<-「そちら」TOS22003-0030-1

C : (ZERO) ditto
C : (NOUN) ditto

D : (PRON) Reference to accumulated contents
TIS12002-330-1|いや|、|今日|の|ところ|は|【これ】|で|い|い|で|す|。

D : (ZERO) ditto
D : (NOUN) ditto

E : (PRON) Difficult to extract antecedent
E : (NOUN) ditto

F : (PRON) Referent is not referred to in the discourse.
F : (ZERO) ditto
F : (NOUN) ditto

G : (PRON) Noun「そのほか」(the other) Case 1
The referent can be explicitly extracted.

G : (NOUN) Generic Use

H : (PRON) Noun「そのほか」(the other) Case 2
The referent cannot be explicitly extracted.

H : (ZERO) Hearer
(NOUN) ditto

[N.B. The meanings of G & H completely differ for NOUN and for the other.]

I : (ZERO) Reference to Unspecified People
(NOUN) Reference to Unspecified Objects

K : (ZERO)「こちら」: The Group/Organization/Place where the speaker belongs.
(NOUN) ditto
(PRON) ditto

N : (NOUN) Proper Name

P : (NOUN) Predicative Use (e.g. It is a *dog.*)

Q : (NOUN) Interrogative or Nouns modified by Interrogative

S : (ZERO) Speaker
(NOUN) ditto

T : (ZERO) Copular Topic Introduction
The OBJECT zero pronoun for topic introduction construction such as 「TOPIC-ですが、....」.
[N.B. Not to be confused with the first character "T" of utterance IDs specifying antecedents.]

V : (NOUN) The antecedent is the nominal form of verbs/adjectives or a <サ変名詞>.

X : (NOUN) Indexical
(PRON) ditto (「こちら」「当」「ここ」)

Y : (ZERO) 「そちら」: The Group/Organization/Place where the speaker belongs.
(NOUN) ditto
(PRON) ditto

a : (PRON) Reference to something in the speaker's mind.
a : (ZERO) ditto
(NOUN) ditto

b : (PRON) Deictic Reference
b : (NOUN) ditto

c : (PRON) Reference to a situation where the dialogue takes place.
c : (ZERO) ditto
(NOUN) ditto

d : (PRON) Meaning "- which you said now" especially the pronominal 「その」 in the sense of "the" (exceptions: 「その後」「その方」「その場合」 etc.).

f : (NOUN) Referent is specified by the modifier

g : (ZERO) Generic Person

n : (NOUN) Introduction of a New Object

p : (NOUN) Adjectival

q : (NOUN) Numerative (e.g., 「ひとり」)

s : (NOUN) Symbol

t : (ZERO) Reference to some intention by a subject OBJE (e.g., given for 「はい、そうです」 or 「その通りです」).

x : (NOUN) Referent difficult to specify (idiomatic expressions or non-verbal abstract nouns)

-： (NOUN) Given for the following cases:

1. The head is 「ほか」 (e.g.,「そのほか」&「（この）ほか」).
2. The following functional nouns with modifiers

「かぎり」：できる【かぎり】のことはいたします。
「かわり」：その【かわり】にシアターバーの方で、ビデオをお楽しみいただけます。
「こと」：いや、来てもらうほどの【こと】でもないので、結構です。
「ため」：移動する【ため】にわざわざ着替えるのもいやですよ。
「つもり」：あさっての二十八日の【つもり】なんですが。
「とおり」：はい、その【とおり】です。
「ところ」：河原町駅からだいたい七、八分といった【ところ】です。
「はず」：電話料金と四百円の手数料の【はず】です。
「ふう」：十四日と十五日、滞在したいという【ふう】に言ってます。
「ほう」：…少々窮屈かと存じますが、ただ広い【方】のツインがふさがっております。
「ほか」：【ほか】に何かございますか。
「まま」：お部屋が少々窮屈かと存じますので、ドアを開けた【まま】にいたしまして…
「まんま」：水温計のレベルは低い【まんま】なんですけどね。
「もの」：月曜日の朝から仕事がある【もの】でして。
「よう」：どうもホテル側のミスの【よう】です。
「わけ」：特にそれにこだわってるっていう【わけ】じゃないんです。

3. The filler「ほう」
4. Non-numerative & non-anaphoric uses of the nouns in 2

?： (ZERO) Invalid zero pronoun (non-obligatory case element)

*： (NOUN) Morpheme analysis error

A2 Statistics of Pronominal Usage

A2.1 Usage of Anaphoric Pronouns

The classification of reference by pronouns in the 375 dialogues which can be used anaphorically (「あそこ」「あちら」「あれ」「いずれ」「おたく」「ここ」「こちら」「こっち」「これ」「そこ」「そちら」「そっち」「それ」「それぞれ」「どちら」「どっち」「どれ」「彼」). (See "Summary of Tags" for the details of the uses.)

Use	Occ.	Description	
T-	658	Anaphoric	
A	42		Reference to Multiple Propositions
C	4		Antecedent within Complex Noun
E	15		Antecedent Difficult to Specify
Y	82	Hearer/Hearer's Place	
K	57	Speaker/Speaker's Place	
X	27	Indexical	
D	12	Accumulative	
F	6	Referent not in Discourse	
a	7	Referent in Speaker's Mind	
b	34	Deictic	
c	7	Reference to Discourse Situation	
d	1	That-Which-You-Said-Right-Now	
Total	952		

A2.2 Usage of Zero Pronouns

The classification of reference by zero pronouns in the 375 dialogues.

A2.2.1 Deep Case[1] (Thematic Role) Distribution

Case	Occ.
AGEN	6810
CONT	6
DEPT	11
DEST	170
EXPR	1380
GOAL	111
IDEN	2
LOCT	60
MUTL	9
OBJE	7531
RECP	1406
RESL	2
ROUT	3
Total	17501

A2.2.2 Referential Properties

Usage	Occ.	Description
T-	3539	Anaphoric
A	262	Antecedents in Multiple Utterances
C	18	Antecedent within Noun Phrase
D	11	Accumulative
F	1745	Referent not in Discourse
H	3038	Hearer
I	96	Unspecified Person
K	1293	Speaker's Place
S	4817	Speaker
Y	481	Hearer's Place
a	10	Referent in Speaker's Mind
c	1	Reference to Discourse Situation
g	422	Generic Person
t	285	Reference to Intention
T	426	Copular Topic Introduction
?	1057	Invalid Zero Nouns
Total	17501	

[1] See TR-IT-0220 for definitions.

A3. Statistics of Pronominal Anaphora

A3.1 Pronoun Antecedent Distance

Dist.	Occ.	Acc.	Acc. %	Non-0 Dist.Acc.	Non-0 Dist.Acc. %
0	113	113	25.17%	-	-
1	230	343	76.39%	230	68.45%
2	61	404	89.98%	291	86.61%
3	20	424	94.43%	311	92.56%
4	8	432	96.21%	319	94.94%
5	4	436	97.10%	323	96.13%
6	3	439	97.77%	326	97.02%
7	2	441	98.22%	328	97.62%
8	1	442	98.44%	329	97.92%
9	3	445	99.11%	332	98.81%
10	1	446	99.33%	333	99.11%
17	1	447	99.55%	334	99.40%
19	1	448	99.78%	335	99.70%
22	1	449	100.00%	336	100.00%

The distance from pronouns to antecedents (only when antecedents are contiguous)

A3.2 Zero Pronoun Antecedent Distance

Dist.	Occ.	Acc.	Acc. %	Non-0 Dist.Acc.	Non-0 Dist.Acc. %
0	460	460	14.73%	-	-
1	940	1400	44.83%	940	35.30%
2	447	1847	59.14%	1387	52.08%
3	280	2127	68.11%	1667	62.60%
4	191	2318	74.22%	1858	69.77%
5	99	2417	77.39%	1957	73.49%
6	91	2508	80.31%	2048	76.91%
7	79	2587	82.84%	2127	79.87%
8	71	2658	85.11%	2198	82.54%
9	51	2709	86.74%	2249	84.45%
10	53	2762	88.44%	2302	86.44%
11	51	2813	90.07%	2353	88.36%
12	38	2851	91.29%	2391	89.79%
13	34	2885	92.38%	2425	91.06%
14	30	2915	93.34%	2455	92.19%
15	18	2933	93.92%	2473	92.87%
16	18	2951	94.49%	2491	93.54%
17	13	2964	94.91%	2504	94.03%
18	14	2978	95.36%	2518	94.56%
19	18	2996	95.93%	2536	95.23%
20	21	3017	96.61%	2557	96.02%
21	17	3034	97.15%	2574	96.66%
22	17	3051	97.69%	2591	97.30%
23	9	3060	97.98%	2600	97.63%
24	13	3073	98.40%	2613	98.12%
25	5	3078	98.56%	2618	98.31%
26	10	3088	98.88%	2628	98.69%
27	4	3092	99.01%	2632	98.84%
28	2	3094	99.07%	2634	98.91%
29	7	3101	99.30%	2641	99.17%
30	4	3105	99.42%	2645	99.32%
31	1	3106	99.46%	2646	99.36%
32	3	3109	99.55%	2649	99.47%
33	2	3111	99.62%	2651	99.55%
35	3	3114	99.71%	2654	99.66%
39	1	3115	99.74%	2655	99.70%
42	1	3116	99.78%	2656	99.74%
43	1	3117	99.81%	2657	99.77%
44	1	3118	99.84%	2658	99.81%
45	1	3119	99.87%	2659	99.85%
46	2	3121	99.94%	2661	99.92%
49	1	3122	99.97%	2662	99.96%
55	1	3123	100.00%	2663	100.00%

The distance from zero pronouns to antecedents
(only when antecedents are contiguous)

A3.3 Pronoun Antecedent Parts-of-Speech

A3.3.1 Overall Statistics

The parts of speech of the antecedents of anaphoric pronouns in the 375 files.

Parts-of-Speech	Occ.	%
<普通名詞>	224	25.99
<固有名詞>	172	19.95
<語尾>	125	14.50
<代名詞>	111	12.88
<名詞句>	103	11.95
<接尾辞>	37	4.29
<サ変名詞>	13	1.51
<格助詞>	11	1.28
<数詞>	9	1.04
<副詞>	8	0.93
<日時>	7	0.81
<終助詞>	7	0.81
<本動詞>	5	0.58
<補助動詞>	5	0.58
<副助詞>	5	0.58
<連体詞>	5	0.58
<人名>	3	0.35
<助動詞>	3	0.35
<接続助詞>	2	0.23
<感動詞>	2	0.23
<並立助詞>	1	0.12
<引用助詞>	1	0.12
<接頭辞>	1	0.12
<助動詞語幹>	1	0.12
<準体助詞>	1	0.12
TOTAL	862	100%

A3.3.2 Breakdown

Nominal

Parts-of-Speech	Occ.	%
<普通名詞>	224	25.99
<固有名詞>	172	19.95
<代名詞>	111	12.88
<名詞句>	103	11.95
<接尾辞>	37	4.29
<サ変名詞>	13	1.51
<数詞>	9	1.04
<副詞>	8	0.93
<日時>	7	0.81
<副助詞>	5	0.58
<人名>	3	0.35
<準体助詞>	1	0.12
TOTAL	693	80.39

Propositional

Parts-of-Speech	Occ.	%
<語尾>	125	14.50
<終助詞>	7	0.81
<本動詞>	5	0.58
<補助動詞>	5	0.58
<助動詞>	3	0.35
<助動詞語幹>	1	0.12
TOTAL	146	16.93

Others (See below for examples)

Parts-of-Speech	Occ.	%
<格助詞>	11	1.28
<連体詞>	5	0.58
<接続助詞>	2	0.23
<感動詞>	2	0.23
<並立助詞>	1	0.12
<引用助詞>	1	0.12
<接頭辞>	1	0.12
TOTAL	23	2.67

Examples of <格助詞> (case marker) Antecedents

Most of them are regarded as propositional.

TBS22003-0250-1 12 (で <格助詞>)

TBS22004-0270-1 13 (で <格助詞>)

「…でよろしいでしょうか。」「それでおねがいします。」

TCS22052-0210-2 14 (で <格助詞>)

「…ことでよろしいでしょうか。それでよろしゅうございますか。」

TCS22061-0140-1 23 (で <格助詞>)
「…ことでよろしいでしょうか。」「それで結構です。」

TCC23054-0050-1 4 (につきまして <格助詞>)
「当日の配達につきましては…」「いつそれがわかるでしょうか。」:F

TCC23054-0070-1 20 (に <格助詞>)
「…という形になるかと思います。」「それでおねがいします。」

TCS32008-0070-1 4 (で <格助詞>)
「…でよろしゅうございますね。」「それで結構です。」

TDS12005-0200-2 21 (で <格助詞>)
「…ことでよろしいですか。」「それで結構です。」

TOS12002-0130-1 4 (にかけて <格助詞>)
「チャイナタウンにかけて…」「そこには…」

TOS32003-0130-3 17 (で <格助詞>)
「お食事ということで。」「それでいかがですか。」:A

TSS12001-0240-2 3 (から <格助詞>)
「七時から、それで結構でございます。」

Examples of <連体詞> (prenominal) Antecedents

TAS12012-0110-1 1 (当 <連体詞>)
「当ホテル」「そちらのホテル」

TAS32004-0230-2 8 (その <連体詞>)
「そのあたり」「こちらのホテル」

TCS13020-0110-3 1 (この <連体詞>)
「このお寺」「こちらの方」

TGS13002-0160-2 1 (この <連体詞>)
「この前(=これの前)」「これでお願いします。」

TOS22003-0240-2 1 (その <連体詞>)
「その場所(=店の場所)」「そこは…」

Examples of <接続助詞> (conjunctive particle) Antecedents

TCS33015-0280-2 16 (て <接続助詞>)
「百メートルほど歩かれまして、そちらの…」

TDS33007-0140-1 8 (ば <接続助詞>)
「お昼までちょうどかかれば、それで結構なんですが。」

Examples of <感動詞> (interjection) Antecedents

TAS13015-0370-1 1 (それは <感動詞>)
「それは良かった。これで…もできますね。」

TAS33015-0110-1 1 (それは <感動詞>)
「それはなかなかよさそうですね。それにしたいなあ。」

Example of <並立助詞> (coordinate particle) Antecedents

TAS33018-0060-2 5 (と <並立助詞>)
「ツインとダブルと、どちらが…」

Example of <引用助詞> (quotative particle) Antecedents

TAS13016-0310-1 27 (かどうか <引用助詞>)
「…かどうか、それを教えていただきたい…」

Antecedents of COND Anaphoric Pronouns

A typical example of COND case pronouns is 「それでけっこうです。」 "*That*'s fine." The COND case pronouns have more propositional antecedents than nominal antecedents (cf. Antecedents of Non-COND Anaphoric Pronouns below).

Propositional	108
Nominal	95
Others	2
Total	225

Antecedents of Non-COND Anaphoric Pronouns

Propositional	45
Nominal	591
Others	9
Total	645

A3.4 Zero Pronoun Antecedent Parts-of-Speech

A3.4.1 Overall Statistics

Parts-of-Speech	Occ.	%
<普通名詞>	1461	41.51
<名詞句>	801	22.76
<語尾>	407	11.56
<サ変名詞>	362	10.28
<固有名詞>	197	5.60
<代名詞>	95	2.70
<接尾辞>	49	1.39
<終助詞>	25	0.71
<格助詞>	22	0.62
<助動詞>	17	0.48
<副助詞>	12	0.34
<人名>	11	0.31
<副詞的名詞>	11	0.31
<本動詞>	11	0.31
<準体助詞>	6	0.17
<引用助詞>	5	0.14
<助動詞語幹>	5	0.14
<日時>	5	0.14
<補助動詞>	5	0.14
<係助詞>	3	0.09
<数詞>	3	0.09
<接続助詞>	2	0.06
<並立助詞>	2	0.06
<感動詞>	1	0.03
<副詞>	1	0.03
<補助動詞語幹>	1	0.03
Total	3520	100%

A3.4.2 Breakdown

Nominal

Parts-of-Speech	Occ.	%
<普通名詞>	1462	41.53
<名詞句>	801	22.76
<サ変名詞>	362	10.28
<固有名詞>	197	5.60
<代名詞>	95	2.70
<接尾辞>	49	1.39
<副助詞>	12	0.34
<人名>	11	0.31
<副詞的名詞>	11	0.31
<準体助詞>	6	0.17
<日時>	5	0.14
<数詞>	3	0.09
<副詞>	1	0.03
Total	3015	85.65

Propositional

Parts-of-Speech	Occ.	%
<語尾>	407	11.56
<終助詞>	25	0.71
<助動詞>	17	0.48
<本動詞>	11	0.31
<助動詞語幹>	5	0.14
<補助動詞>	5	0.14
<補助動詞語幹>	1	0.03
Total	471	13.38

Others (See below for examples)

Parts-of-Speech	Occ.	%
<格助詞>	22	0.62
<引用助詞>	5	0.14
<係助詞>	3	0.09
<接続助詞>	2	0.06
<並立助詞>	2	0.06
Total	34	0.97

Examples of <格助詞> (case marker) Antecedents

Only one case is inter-sentential anaphora

```
[TDS33008-0280-1]
1              2  3  4      5  6  7  8  9     10 11 12     13 14 15 16 17
お#座敷#てんぷら で 、 御#予算 は お 一 人 一#万 円 で よろし い で す ね 。
[TDS33008-0290-1]
1   2  3   4   5  6  7
ええ 、 大変 結構 で す 。
  結構 (4)
    (OBJE NIL)   TDS33008-0280-1,1-11
```

The rest are actually subsumption rather than anaphora. In these cases, the obligatory cases were substituted for by the case elements tagged as the antecedents.

```
[TCC22033-0050-2]
1  2  3   4   5  6  7    8  9   10 11 12 13 14 15 16 17 18 19 20
先 に 荷物 だけ 運 ん でもら う よう に お 願 い し た ん で す が 。
  願 (12)
    (OBJE NIL)   TCC22033-0050-2,1-10
```

(Here, <格助詞> "に" (10) has been assigned no deep case in the case analysis database.)

Example of <引用助詞> (quotative particle) Antecedents

In all cases, <引用助詞> was "か" (if) or "かどうか" (whether), asking the truth values of propositions.

```
[THS32002-0040-2]
1      2   3  4        5  6  7  8  9    10      11 12
それぞれ 何時 に ワシントン に 着 く か 教え てくださ い 。
[THS32002-0050-1]
1  2   3  4  5  6    7  8   9  10 11 12    13 14
お 調べ し ま す ので 、 少し お 待 ち くださ い 。
  調べ (2)
    (OBJE NIL)   THS32002-0040-2,1-8
```

Examples of <係助詞> (topic marker) Antecedents

There were two cases of cataphora, one of which is

```
[TSS32001-0160-2]
1         2     3  4   5  6
それでしたら いかが で しょ う 。
  で (3)
    (OBJE NIL)   TSS32001-0160-3,1-25
[TSS32001-0160-3]
1   2   3  4   5  6  7  8   9  10  11 12 13 14    15  16     17 18
少し 早め に バー の 方 に お越 し にな っ て 、 食前酒 でも 召し上が り ながら
     19  20 21  22 23 24 25 26
     夜景 を 楽し ま れ て は 。
```

There was also a long distance anaphora whose antecedent is the topic of the entire discourse:

```
[TOS32006-0060-1]
  1                  2        3  4  5      6      7      8  9  10 11 12 13 14 15
  ヘリコプターツアー っていう の は いくつ ぐらい コース が あ る ん で す か 。
   ......
[TOS32006-0220-1]
  1    2  3  4  5  6      7    8    9        10   11 12     13 14 15 16 17 18 19 20 21 22
  そう で す か 、 じゃあ ぜひ その 八#十#五 ドル の コース に し た い と 思 い ま す 。
    し (14)
      (OBJE NIL)    TOS32006-0060-1,4
```

Examples of <接続助詞> (conjunctive particle) Antecedents

Two cases where <接続助詞> "て" has the OBJE case

```
[TAS32004-0170-2]
1    2  3        4  5    6  7  8
では お 役に立て て 光栄 で す 。
  光栄 (5)
    (OBJE NIL)    TAS32004-0170-2,2-4
[TCS12005-0270-2]
1  2  3  4  5    6  7  8  9
近 く て 良 かっ た で す 。
  良 (4)
    (OBJE NIL)    TCS12005-0270-2,1-3
```

Examples of <並立助詞> (exemplar particle) Antecedents

The <並立助詞> "とか" is used to exemplify nominals.

```
[TCC22074-0040-1]
1    2  3  4  5      6    7        8    9    10 11 12 13 14 15       16
でき れ ば 、 体温計 とか 氷まくら とか 借り た い ん で す けれども 。
  借り (9)
    (OBJE NIL)    TCC22074-0040-1,5-8
```

The <並立助詞> "たり" is used to exemplify verbals.

```
[TCS13019-0160-1]
1  2  3  4          5  6  7          8  9  10   11 12     13 14 15 16   17
朝 は 、 お＃泊まり の 方 全員＃参加 で 、 住職 の お＃話 を 聞 い たり 、
          18       19 20   21 22   23 24   25   26 27 28 29
          それから 、 座禅 を 体験 し たり でき ま す よ 。
  でき (25)
    (OBJE NIL)    TCS13019-0160-1,10-24
```

A3.5 Regular Pronoun Antecedent Usage

A3.5.1 The Usage of Nouns as Antecedents of Pronouns

The usage of the last morphemes of antecedents (see Part I for the definition of usage). The Rate is the occurrence of the use as antecedent divided by the total occurrence of the use, which represents the likelihood for uses to be antecedents.. For example, 0.3% of X (indexical) became antecedents of some pronouns.

Usage	Rate	%	Description
I	121/3615	3.3%	Non-specific
f	44/1841	2.4%	Definite by Modifier
F	38/1901	2.0%	Definite by Background
T/C	32/1622	2.0%	Anaphoric
N	29/233	12.4%	Proper Noun
n	23/239	9.6%	New Object
P	18/465	3.9%	Predicative
-	7/1269	0.6%	Complementizer
p	4/1139	0.4%	Adjectival
q	4/178	2.3%	Numerative
s	3/151	2.0%	Symbolic
V	2/699	0.3%	Verbal
D	1/11	9.1%	Accumulated Contents
b	1/15	6.7%	Deictic
X	1/343	0.3%	Indexical
Avr.	328/14669	2.2%	Average

Note Although one might expect that nouns referring to non-specific objects (e.g., I & P) are less likely to be antecedents, their rates of being antecedents are higher than those of the uses f, F & T which refer to specific objects. (This may be partly because of the nature of the dialogues (i.e., the reservation task)).

Some peculiar examples of antecedents

- Complementizer

TAC23023-0150-2,19

「ツインルームということで」<-「それで結構です。」

TAS13001-0320-2,9

「朝食は洋食ということで」<-「それでお願いします。」

TAS22031-0200-1,21)

「エキストラベッドをお入れすることもできます」<-「それでお願いします。」

TAS32012-0180-1,11

「三泊の予定ということで」<-「それで結構です。」

TAS33017-0270-1,11

「キャンセル待ちということで」<-「それで結構です。」

TGS22001-0150-1,12

「キャンセルなさるということですが、それでよろしいでしょうか。」

TCC22013-0050-2,5

「今すぐチェックインすることは」<-「それは無理なようなんです。」

p Adjectival

TAC22012-0180-1,8

「満室」<-「どちらにしても」（満室であろうとなかろうと）

TDS33008-0100-1,8

「宴会＃スタイル」<-「それもよさそうですね。」

TOS13003-0180-2,1

「午前のコース」<-「こちらのコース」

q Numerative

TAS12022-0070-1,11

「ツインのお部屋を二部屋」

TDS12008-0070-1,16

「パスタの中からお好きなものを二種類」

TDS12008-0070-1,16

「お好きなパスタを二種類」

THS12002-0190-1,6

「片道＃百＃十＃五＃ドル」<-「これは税込みの価格です。」

s Symbols

TAS32010-0120-1,1； TAS32011-0220-1,9； TAS13015-0270-1,3

電話番号／カード番号 <- 「これ（それ）でよろしいでしょうか。」

V Verbal

TCS13020-0140-1,3； TCS13020-0130-1,3

「お＃勤め」＝「ミサのようなもの」

D Accumulated Contents

TCS22051-0140-2,1

「以上でよろしいでしょうか。」「それで結構です。」

A3.5.2 The Usage of Pronouns as Antecedents of Pronouns

Most of the pronouns which are the antecedents of some pronouns are themselves anaphoric.

Usage	Rate	%	Description
T-	64/725	8.83%	Anaphoric
X	1/27	3.70%	Indexical
D	1/12	8.33%	Accumulated Contents
Avr.	66/1715	3.76%	Average

A3.6 Grammatical Functions of the Antecedents of Regular Pronouns

The grammatical function of the last morphemes of antecedents of pronouns. The rate is the occurrence of the grammatical function as antecedent divided by the total occurrence of the grammatical function, which represents the likelihood for grammatical functions to be antecedents. For example, 33.0% of the が-case became antecedents of some pronouns. Note that these figures may be related to the Forward Center (Cf) priority.

Grammatical Func.	Rate	%
<係助詞>	74/211	35.1%
が	62/191	32.5%
を	22/102	22.6%
に	44/210	21.0%
Other <格助詞>	78/264	29.5%
Other	169/876	19.3%
Average	449/1854	24.2%

A3.7 Grammatical Functions of the Antecedents of Zero Pronouns

The figures can be interpreted similarly with those of the Grammatical Functions of the Antecedents of Regular Pronouns. Note that not much regularity is found between the two tables (A3.6 & A3.7).

Grammatical Func.	Rate	%
<係助詞>	642/3673	17.5%
が	289/2472	11.7%
を	604/2598	23.3%
に	193/4707	4.1%
Other <格助詞>	262/7845	3.3%
Other	1133/19866	5.7%
Average	3123/41161	7.6%

A4 Examination of Errors

This section presents errors found in the best score case of the pronoun anaphora resolution experiment (the Way A of the cross validation).

Format:

```
# <Pronoun Utterance ID>,<Pronoun Morpheme Index>     <Pronoun Case Frame>
{→<Antecedent Candidate Utterance ID>,<Candidate Morpheme Index>   <Candidate Morpheme>}*
<Text> (Pronouns are indicated by bold+italics and antecedents by underlines.)
<Comments>
```

A4.1 Positive Cases: (The system judges antecedents to be non-antecedents.)

1) TAS12016-0110-1,5 ((こちら (の <連体助詞>) (名前 <普通名詞>))

→TAS12016-0100-1,6 (様 <接尾辞>)

担当者：はい、トーマス・ネルソン様、十月二十八日、二十九日のご予約ですね。はい、で、
こちらのお名前をどなたにご変更なさるんでしょうか。

Failed in the Zero Pronoun-Antecedent Combination Check.
Insufficient Training Data

2) TAS32014-0230-1,2 ((それ (まで <格助詞>) (着 <本動詞> 五段カ)) (cf. 14)

→TAS32014-0220-1,11 (六時 <日時>)

担当者：三時にご到着なさいましたら、六時の、サーカスのショーをご覧いただけます。
申込者：じゃあ***それ***までに着くようにします。どうもありがとう。

Failed in the Case-Frame Check (Case-Based) (insufficient <日時>+まで combination in the training data).
Insufficient Training Data

3) TCC22014-0140-1,8 ((こちら (の <連体助詞>) (部屋 <普通名詞>)) (cf. 16)

→TCC22014-0130-1,9 (ほう <普通名詞>)

担当者：デラックスツインルームならございますが、こちらのほうは、すいませんが、二百ドル、一泊二百ドルいたします。もしお客様さえよろしければ、**こちら**のお部屋でしたら今すぐチェックインしていただけますが。

Failed in the Zero Pronoun-Antecedent Combination Check.
Insufficient Training Data

4) TCS13001-0110-1,1 ((こちら (は <係助詞>) (な <本動詞> 五段ラ))

→TCS13001-0100-1,16 (サービス <サ変名詞>)

担当者：そういたしますと翌朝八時までにお部屋にお届けするサービスが有ります。
こちらは普通のクリーニング料金の三十パーセント増しになってしまいますが。

Failed in the Zero Pronoun-Antecedent Combination Check (こちら→<サ変名詞>).

Probably inevitable with the scheme of the experiment.

A4.2 Negative Cases: (The system judges non-antecedent candidates to be antecedents.)

5) TAS12018-0190-2,3 ((それ (で <格助詞>) (願 <本動詞> 五段ワ))

→TAS12018-0150-1,9 (分 <接尾辞>)
→TAS12018-0160-1,8 (追加 <サ変名詞>)
→TAS12018-0160-1,6 (分 <接尾辞>)
→TAS12018-0180-1,18 (こと <普通名詞>)
→TAS12018-0180-1,5 (当り <接尾辞>)

担当者：スイートのお部屋の、二名様分の基本料金が百二十ドルになっております。
で、二名様分の追加ですので、お一人当り十ドルの追加料金が付きます。
ですので、四名様で、一泊当り百四十ドルになります。
お一人様当りに直しますと、一泊三十五ドルということです。
申込者：はい、分かりました。じゃあ、***それ***で予約をお願いします。

"それ" in the COND case can be taken as referring to the entire condition offered.

6) TAS13017-0120-1,3 ((こちら (の <連体助詞>) (パーティプラン <普通名詞>))

→TAS13017-0090-1,14 (円 <接尾辞>)
→TAS13017-0090-1,10 (泊 <接尾辞>)

担当者：はい、こちらはお一人様一泊税込みで一万五千円となっております。
申込者：そうですか。どうもありがとうございました。
担当者：はい、***こちら***のパーティプランは、十一月末までになっております。

Insufficient training data (genitive data tends to be sparse).

7) TAS32004-0160-1,1 ((それ (で <格助詞>) (願 <本動詞> 五段ワ))

→TAS32004-0150-1,29 (方 <普通名詞>)

担当者：お部屋とは別に、税金が十八点二五パーセント、そしてサービスチャージが二ドルかかってまいりますけれども、そちらの、方でよろしいでしょうか。
申込者：***それ***でお願いします。

The result can be taken to be correct.

8) TAS32008-0130-1,7 ((それ (に <格助詞>) (し <本動詞> サ変))

→TAS32008-0120-2,6 (空き <普通名詞>)

申込者：ではダブルの二部屋続きはありますか。
担当者：はい、ございます。ちょうど九月の二十日、空きがございます。
申込者：そうですか、では***それ***にしようかな。

The result can be taken to be correct.

9) TAS32008-0180-1,3 ((そちら (の <連体助詞>) (電話番号 <普通名詞>))

→TAS32008-0170-1,8 (室 <接尾辞>)

担当者：はい、鈴木様、ニューオータニホテルの六百二号室に御滞在中ですね。

そして、**そちら**のお電話番号が、七一四、四四三、一七零零でございますね。

Inevitable with the scheme of the experiment. (The recognizer would have to know that a hotel room normally does not have a direct phone number.)

10) TAS32008-0190-1,3 ((それ (で <格助詞>) (い <形容詞>))

→TAS32008-0180-1,8 (零 <数詞>)

担当者：はい、鈴木様、ニューオータニホテルの六百二号室に御滞在中ですね。
そして、そちらのお電話番号が、七一四、四四三、一七零零でございますね。
申込者：宿泊先は**それ**でいいんですが、電話番号は、二一三、四四三、一七零零です。

Inevitable with the scheme of the experiment. (The recognizer would have to recognize the parallelism in the utterances.)

11) TAS32014-0230-1,2 ((それ (まで <格助詞>) (着 <本動詞> 五段カ))

→TAS32014-0220-1,16 (ショー <普通名詞>)
→TAS32014-0220-1,14 (サーカス <普通名詞>)

担当者：三時にご到着なさいましたら、六時のサーカスのショーをご覧いただけます。
申込者：じゃあ**それ**までに着くようにします。どうもありがとう。

Requires more analysis: the pattern「それまでに」refers to time.

12) TAS33013-0150-1,1 ((そちら (の <連体助詞>) (会社 <普通名詞>))

→TAS33013-0140-1,9 (訪問 <サ変名詞>)

申込者：それと京都では、京都シーエービー社、を訪問しようと思っています。
そちらの会社の方の電話がよろしいですか。

Insufficient Training Data

13) TBS32004-0150-1,1 ((こちら (で <格助詞>) (よろし <形容詞>))

→TBS32004-0140-1,13 (ドル <接尾辞>)
→TBS32004-0140-1,8 (ダブルルーム <普通名詞>)
→TBS32004-0140-1,6 (ドル <接尾辞>)
→TBS32004-0140-1,1 (シングルルーム <普通名詞>)

担当者：お待たせしました。マジソンスクエアガーデンの近くのホテルニューヨークがお取りできます。シングルルームが一泊百三十ドル、ダブルルームが一泊百八十ドルとなっております。**こちら**でよろしいでしょうか。

Inevitable with the scheme of the experiment. (The recognizer would have to know the task structure.)

14) TBS33001-0260-1,1 ((こちら (は <係助詞>) (付 <本動詞> 五段カ))

→TBS33001-0250-1,5 (八 <数詞>)
→TBS33001-0250-1,3 (二 <数詞>)

担当者：はい、そうでございます。二つ目が同じく祇園にあります加茂川ホテルでございます。電

話番号が零七五、四四二の一零八八です。**こちら**は食事は付いてないんですが、一泊二万円でございます。

Insufficient Training Data ?

15) TCC22011-0090-1,1 ((それ (は <係助詞>) (ぐらい <副助詞>))
→TCC22011-0080-2,9 (用意 <サ変名詞>)
→TCC22011-0080-2,7 (しか <副助詞>)

担当者：はい。今でしたら、デラックスツインしかご用意できないんですが。
申込者：***それ***は一泊おいくらぐらいですか。

Insufficient Training Data ?

16) TCC22014-0140-1,8 ((こちら (の <連体助詞>) (部屋 <普通名詞>))
→TCC22014-0130-1,20 (ドル <接尾辞>)
→TCC22014-0130-1,15 (ドル <接尾辞>)

担当者：デラックスツインルームならございますが、こちらのほうは、すいませんが、二百ドル、一泊二百ドルいたします。もしお客様さえよろしければ、**こちら**のお部屋でしたら今すぐチェックインしていただけますが。

Insufficient Training Data ?

17) TCC22071-0140-1,6 ((それ (は <係助詞>) (もの <普通名詞>))
→TCC22071-0130-1,10 (ほう <普通名詞>)
→TCC22071-0130-2,1 (何 <代名詞>)

担当者：すみませんが、おかゆというもの、ちょっとわたくしのほうでは知らないんですが。何ですか。
申込者：そうですか、***それ***は、普通にごはんを炊くよりも多めに水を入れて炊いたものなんです。

Inevitable with the scheme of the experiment. (The recognizer would have to recognize the parallelism in the utterances.)

18) TCC22071-0150-1,15 ((それ (を <格助詞>) (作れ <本動詞> 一段))
→TCC22071-0140-1,19 (水 <普通名詞>)
→TCC22071-0140-1,11 (ごはん <普通名詞>)

申込者：そうですか、それは、普通にごはんを炊くよりも多めに水を入れて炊いたものなんです。
担当者：はい、分かりましたが、料理長のほうがちょっと***それ***を作れるかどうかは、はっきりここで申し上げられません。

Inevitable with the scheme of the experiment. (cf.「それ」 here refers to the topic.)

19) TCC22071-0160-1,16 ((それ (が <格助詞>) (作れ <本動詞> 一段))
→TCC22071-0150-1,23 (ここ <代名詞>)

担当者：はい、分かりましたが、料理長のほうがちょっとそれを作れるかどうかは、はっきりここで申し上げられません。それではですね、ちゅう房のほうに参りまして、***それ***が作れるか

どうか、聞いて参ります。

Insufficient Training Data (cf.「それ」 here refers to the topic.)

20) TCS13001-0120-2,1 ((それ (で <格助詞>) (願 <本動詞> 五段ワ))

→TCS13001-0110-1,3 (普通 <普通名詞>)

→TCS13001-0120-1,3 (それ <代名詞>)

担当者:こちらは普通のクリーニング料金の三十パーセント増しになってしまいますが。

申込者:ええ、それで結構ですよ。***それ***でお願いします。

Inevitable with the scheme of the experiment (a difficult case).

21) TCS22022-0160-1,2 ((こちら (の <連体助詞>) (ほう <普通名詞>))

→TCS22022-0150-1,25 (会社 <普通名詞>)

→TCS22022-0150-1,20 (ベビーシッター <普通名詞>)

→TCS22022-0150-1,14 (資格 <普通名詞>)

担当者:もう一つのほうの会社ですけれども、こちらは資格を持たないベビーシッターを抱えている会社です。普通**こちら**のほうは特別なケースのために教育を受けた人はいないと思うんですが。

Inevitable with the scheme of the experiment (a first wrong record may be saved by the analysis of the copula pattern).

22) TCS22022-0180-1,3 ((そちら NIL (願 <本動詞> 五段ワ))

→TCS22022-0170-1,20 (説明 <サ変名詞>)

→TCS22022-0170-1,17 (こと <普通名詞>)

担当者:そうしましたら、資格を持った専門のベビーシッターを抱えている会社のことについてご説明いたしましょうか。

申込者:はい、**そちら**お願いします。

Inevitable with the scheme of the experiment (the Zero Pronoun-Antecedent Combination Check cannot screen the (そちら <サ変名詞>) combination, for そちら may refer to <サ変名詞>).

23) TCS22022-0240-1,1 ((こちら (の <連体助詞>) (ベビーシッター <普通名詞>))

→TCS22022-0200-1,4 (年齢 <普通名詞>)

→TCS22022-0210-1,15 (障害 <普通名詞>)

→TCS22022-0210-1,13 (体 <普通名詞>)

→TCS22022-0210-1,3 (おしめ <普通名詞>)

→TCS22022-0220-1,6 (前 <接尾辞>)

担当者:はい、分かりました。会社の名前はナースリーライムと申します。基本的にどんな年齢のお子様でもお世話できます。例えば、おしめをしていらっしゃるお子様ですとか、体に障害をお持ちのお子様などですね。ですが、通常一日前にはご予約いただくことが必要となっております。また、続けて何日かサービスを提供するのはちょっと難しいか

と思うんですが。***こちら***のベビーシッターは人命救助の免許を持っております。

Insufficient Training Data ?

24) TGS12001-0160-1,1　　　**((それ (は <係助詞>) (有 <本動詞> 五段ラ))**

→TGS12001-0140-1,9　(発 <接尾辞>)
→TGS12001-0140-1,7　(ジェーエフケー空港 <固有名詞>)
→TGS12001-0150-1,3　(着 <接尾辞>)
→TGS12001-0150-1,1　(エルエーエックス <固有名詞>)

担当者:この便ですと、ジェーエフケー空港を十五時発になります。エルエーエックスに十七時三十分着となります。
申込者:***それ***は空席は有りますか。

Inevitable with the scheme of the experiment.　(The recognizer would have to know that 便 (flights) have seats.)

25) TGS12001-0180-1,3　　　**((それ (を <格助詞>) (予約 <サ変名詞>))**

→TGS12001-0160-1,3　(空席 <普通名詞>)
→TGS12001-0170-1,4　(空席 <普通名詞>)

申込者:それは空席は有りますか。
担当者:はい、まだ空席がございます。
申込者:じゃ、***それ***を予約します。

Inevitable with the scheme of the experiment?　(An vacant seat may be reservable, but there is no training data for reserving vacant seats.)

26) TKS13002-0140-2,1　　　**((それ (で <格助詞>) (願 <本動詞> 五段ワ))**

→TKS13002-0110-1,23　(オートマチック <普通名詞>)

担当者:はい、日曜日の十時からお貸しできるのはシビックなんですけども、それはオートマチックです。
申込者:それから、保険についてはどうしたらいいんですか。
担当者:料金の六千九百円の中に免責補償料も含まれております。
申込者:分かりました。***それ***でお願いします。

Inevitable with the scheme of the experiment (may be saved by the analysis of the copula pattern).

27) TOS23003-0150-1,1　　　**((あそこ NIL (だ <助動詞>))**

→TOS23003-0130-1,7　(寺 <普通名詞>)

申込者:金閣寺というのは金色の寺ですか。
担当者:はい、さようでございますが。
申込者:***あそこ***なら行ったことがあります。

Inevitable with the scheme of the experiment (may be saved by the analysis of the copula pattern).

28) TOS32001-0100-1,1 ((そこ (まで <格助詞>) (いらっしゃ <本動詞> 特殊ラ))
→TOS32001-0090-1,5 (北側 <普通名詞>)

担当者:近代美術館はロックフェラーセンターの一ブロック北側に行ったところです。
そこまでいらっしゃればすぐにわかると思いますよ。

This is not necessarily wrong.

29) TOS32002-0110-1,2 ((そこ (まで <格助詞>) (いらっしゃ <本動詞> 特殊ラ))
→TOS32002-0100-1,5 (北側 <普通名詞>)

担当者:美術館はロックフェラーセンターから一ブロック北側に行ったところにあります。
ですからそこまでいらっしゃれば美術館はすぐにおわかりになると思いますよ。

This is not necessarily wrong.

30) TOS33002-0150-1,1 ((そちら (の <連体助詞>) (方 <普通名詞>))
→TOS33002-0140-1,16 (運行 <サ変名詞>)

担当者:そうですか、私どもも庭園だけめぐるコースをいくつか運行いたしておりますが。
そちらの方がきっとお気に召すかと思います。

Requires more analysis: the recognizer could exclude [動詞 +] <サ変名詞> from candidates.

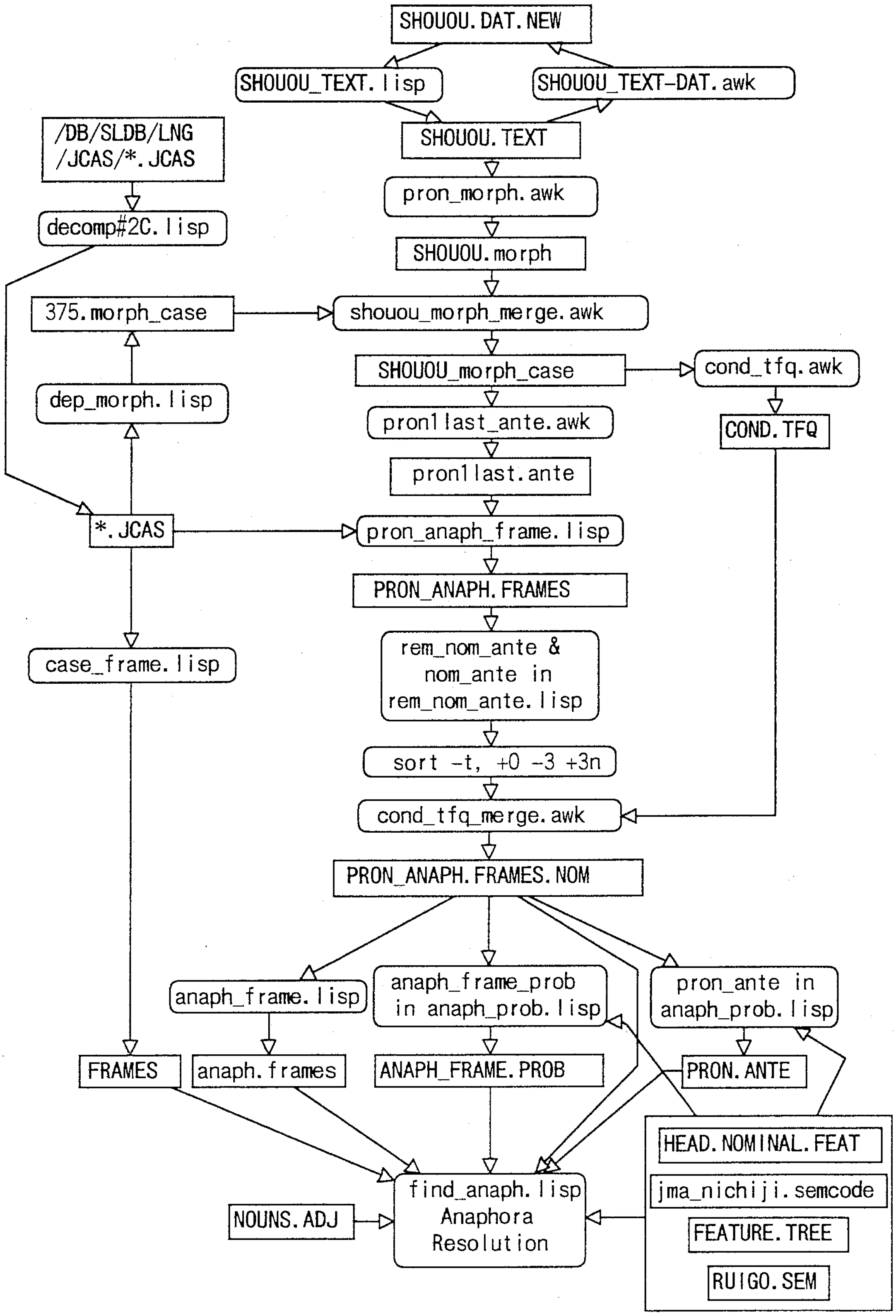
Summary of the causes of the failures

Inevitable: 13
Insufficient Training Data: 10
Requires more analysis: 2
Possibly wrong antecedent: 5

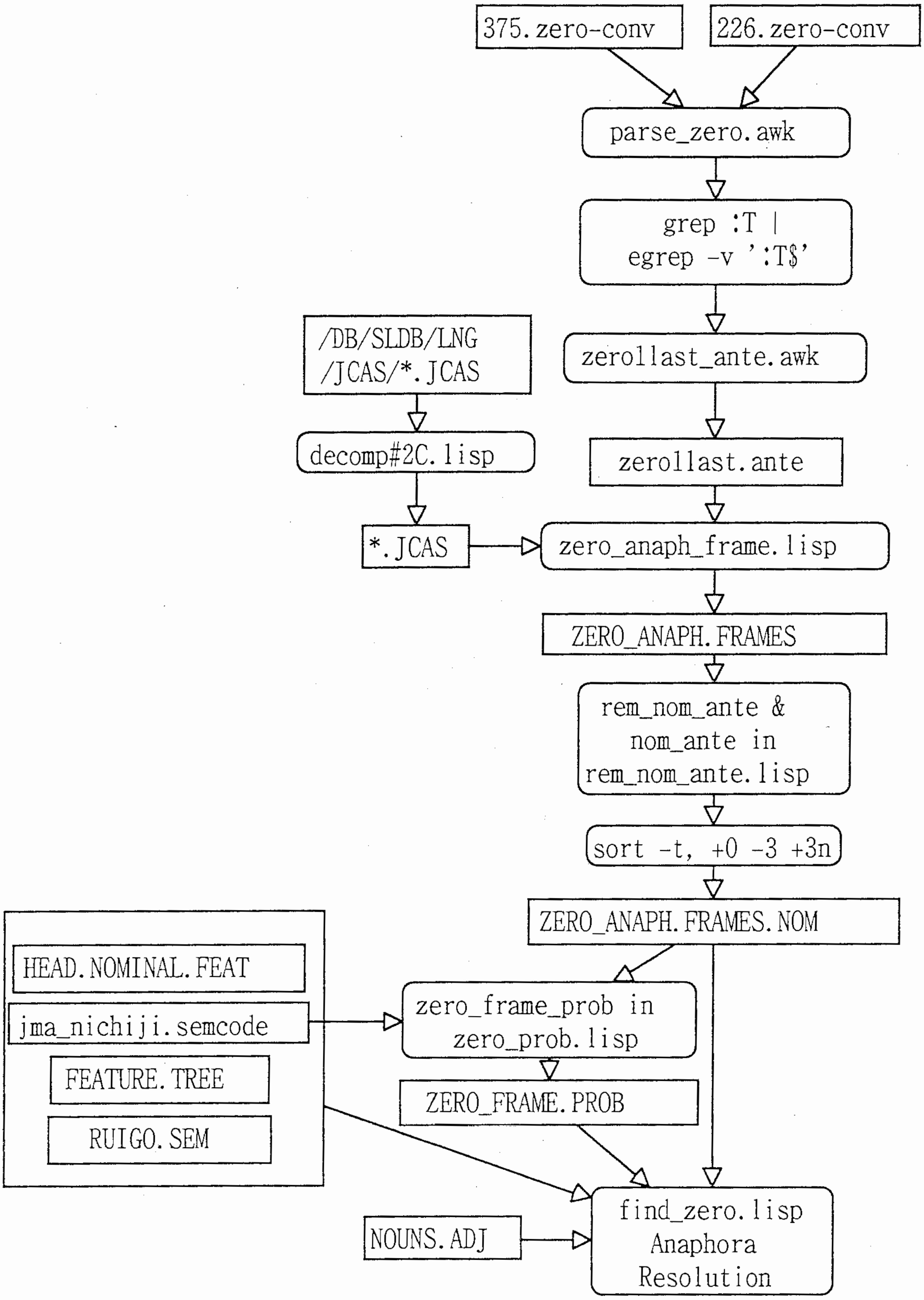
The same pronouns in 2 and 14 and 3 and16 appear in both negative and positive examples, and they are failed for Insufficient Training Data. When the duplication and the cases with possibly wrong antecedents, we have 23 cases. Now, inevitably failed cases with the current scheme are 11 among them. Then, the success rate of the current method in Way A with completed data and analysis would be 82.5% (52/63; 63=68-5 (# of possibly wrong antecedents)).

A5 Data Flow Diagrams for Pronoun Anaphora Resolution Experiments

A5.1 For Regular Pronouns



A5.2 For Zero Pronouns



A5.3 Index for the Data Flow Diagrams

Programs (in rounded boxes)

anaph_frame.lisp	Makes stat files for find_anaph
anaph_prob.lisp	Makes stat files for find_anaph
case_frame.lisp	Extracts case frames from *.JCAS
cond_tfq.awk*	Extracts COND from the prev. utterance
cond_tfq_merge.awk*	Adds COND info to PRON_ANAPH.FRAMES
decomp#2C.lisp	Reformats *.JCAS
dep_morph.lisp	Dumps the morphemes in *.JCAS with case info
find_anaph.lisp	Pronoun Anaphora Resolution
find_zero.lisp	Zero-Pronoun Anaphora Resolution
parse_zero.awk	Reformats *.zero-conv
pron1last_ante.awk*	Extracts pronouns with single antecedents
pron_anaph_frame.lisp	Makes a pronoun-antecedent candidate pair file
pron_morph.awk	Extracts pronouns from SHOUOU.TEXT
rem_nom_ante.lisp	Remove nominals from pron-antecedent files
shouou_morph_merge.awk	Merges morph files
SHOUOU_TEXT.lisp	Reformats SHOUOU.DAT.NEW into SHOUOU.TEXT
SHOUOU_TEXT-DAT.awk	Reformats SHOUOU.TEXT back into SHOUOU.DAT.NEW
zero1last_ante.awk*	Extracts zero pronouns with single antecedents
zero_anaph_frame.lisp	Makes a zero pronoun-antecedent candidate pair file
zero_prob.lisp	Makes stat files for find_zero

Data (in rectangular boxes)

*.JCAS	Case analysis files
*.zero-conv	Zero pronoun usage original data
375.morph_case	Morphemes with their deep cases
anaph.frames	Case frames from PRON_ANAPH.FRAMES.NOM
ANAPH_FRAME.PROB	Case frame with probabilities by pos/neg cases
FEATURE.TREE	Feature inheritance table
FRAMES	Case frames from *.JCAS
HEAD.NOMINAL.FEAT	Feature dictionary
jma_nichiji.semcode	Semcode dictionary
NOUN.ADJ	Nouns in modifier use
pron1last.ante	Pronouns with single antecedents
PRON_ANAPH.FRAMES	Pronoun-antecedent candidate pairs
PRON_ANAPH.FRAMES.NOM	Pronoun-antecedent candidate pairs without Nominals
PRON.ANTE	Pronoun antecedent collocation
RUIGO.SEM	Semcode-Feature table
SHOUOU.DAT.NEW	Pronominal anaphora original data
SHOUOU.TEXT	Pronominal anaphora data reformatted for readability
SHOUOU.morph	Pronouns in SHOUOU.TEXT
SHOUOU_morph_case	Morphemes with anaphora info
ZERO_ANAPH.FRAMES	Zero pronoun-antecedent candidate pairs
ZERO_ANAPH.FRAMES.NOM	Zero pronoun-antecedent candidate pairs without Nominals
ZERO_FRAME.PROB	Stat file for zero pronoun anaphora resolution experiments
zero1last.ante	Zero pronouns with single antecedents

A6 Location of Data

Noun Usage

/dept4/work7/IR_REF/REF9709/*.hutu*-conv

Case Analysis

/DB/SLDB/LNG/JCAS/T*.JCAS

cf. The files were converted with decomp#2C.lisp for the experiments.

Pronoun Usage and Anaphora

/dept4/work7/IR_REF/REF9709/SHOUOU.DAT.NEW

Zero Pronoun Usage and Anaphora

/dept4/work7/IR_REF/REF9709/{226,375-226}.zero-conv

Semcodes

The original	/usr/local/TDMT/tdmt-multi-dev/j-morph/dic/jma-atr-sem-code.text
The one used	/dept4/work7/ANAPH/DATA/jma_nichiji.semcode (<日時> entries added)

Semantic Features

The Semantic Feature Dictionary

/dept4/work7/ANAPH/DATA/375.HEAD.NOMINAL.FEAT

Semcode-Semantic Feature Table

/dept4/work2/PS_WORK/DATA/RUIGO/RUIGO.SEM

Semantic Feature Inheritance Table

/dept4/work2/PS_WORK/DATA/RUIGO/FEATURE.TREE

Nouns of Modifier Use

/dept4/work7/ANAPH/DATA/375.NOUNS.ADJ

A7 Location of Programs

/dept4/work7/ANAPH/prog/

/dept4/work7/IR_IFT/PRON_NF/

```
dep_morph.lisp
pron_morph.awk
SHOUOU_TEXT.lisp
SHOUOU_TEXT-DAT.awk
```