gefördert vom Bundesministerium für Forschung und Technologie

Verbmobil
Verbundvorhaben

Zero pronoun processing: some requirements for a VERBMOBIL system.

Dieter Metzing
Melanie Siegel

Universität Bielefeld

Memo 46
September 1994

September 1994

Dieter Metzing
Melanie Siegel

Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft
– Computerlinguistik –
Pf 100 131
D 33 502 Bielefeld

Tel.: (0521) 106 - 2996
Fax: (0521) 106 - 2996
e-mail: SIEGEL@LILI3.UNI-BIELEFELD.DE

Gehört zum Antragsabschnitt: 12 Transfer(Japanisch)

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 101 G gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Contents

Abstract

Some requirements for a VERBMOBIL system capable of processing Japanese dialogue input have been explored. Based on a pilot study in the VERBMOBIL domain, dialogues between 2 participants and a professional Japanese interpreter have been analyzed with respect to a very typical and frequent feature: zero pronouns. Zero pronouns in Japanese texts or dialogues as well as overt pronouns in English texts or dialogues are an important element of discourse coherence. As to translation, this difference in the use of pronouns is a case of translation mismatch: information not explicitly expressed in the source language is needed in the target language. (Verb argument positions, normally obligatory in English, are rather frequently omitted in Japanese. Furthermore, verbs in Japanese are not marked with respect to features necessary for pronoun selection in English.)

Though zero pronouns have been a prominent topic of linguistic or computational linguistic studies, they have hardly been studied with respect to dialogues occurring in a specific task-domain or with respect to translation procedures in such a domain. An analysis of our data leads to the conclusion that general interpretation procedures for zero pronouns in Japanese that have been developed in discourse analysis or text processing are less applicable to dialogues of a task-specific domain (cf. Siegel & Metzing 1994). Therefore, domain-specific interpretation procedures are required. Two types of procedures are proposed: 1. procedures that process lexical default information, e. g. by default unification. 2. special semantic interpretation procedures that process domain knowledge (task-specific or dialogue structure specific), formally to be described e. g. by circumscription.

1 Zero pronouns: a case study

One issue in the development of translation systems for spontaneous dialogues is the 'real data' issue. Real data, on different levels of analysis, may mean: dictionary and corpus research; research on robust processing to handle non-standard aspects of utterances; research on specific phenomena, case studies, as for example: zero pronouns in specific task-oriented dialogues.

Real data may be used in different ways in the development of processing systems. They play a role in the development of components and tools, or in the definition of 'realistic data', data of reduced complexity that can still be processed by available methods. Properties of such realistic data will then set the stage for the development and evaluation of processing systems. (E. g. the processing of different types of pronouns has been a kind of benchmark test for natural language front ends or dialogue systems.) Since the processing of zero pronouns is a rather complex subject, it is important to determine 'realistic cases', that are still solvable in a formal way and that do not require too much processing resources.

The topic of our case study has been: the *use* of zero pronouns in scheduling dialogues, between 2 participants whose contributions were translated by a professional (Japanese) interpreter (Siegel 1993; Siegel et al. 1993), and *requirements* for a dialoge translation system like VERBMOBIL.

Our case study contributes to the following research questions:

- (a) How complex is the use of zero pronouns in a scheduling domain?
- (b) Which subtypes of zero pronouns appear in which frequency?
- (c) What is the role of text semantic regularities and of domain knowledge for an interpretation of zero pronouns?
- (d) Is there a preferred order in the application of different interpretation strategies? (cf. c)
- (e) To which extend may interpretation requirements of zero pronouns be used to delimit the amount of background knowledge (structure and content) and to develop different layers of a dialogue model?

Zero pronouns are special in the sense that:

- the speech component can contribute no information to their interpretation;
- languages may use different discourse context conventions (overt pronouns/zero pronouns).

- (f) In which sense does the difference between 'overt pronoun language' and 'zero pronoun language' confirm/disconfirm the generality of a component of a processing system or representations, processing procedures, strategies?

2 Zero pronouns in Japanese discourse

In Japanese nominals expressing major grammatical relations such as the subject and object of a sentence can simply be omitted (Kameyama 1985; Kameyama 1989). Zero pronouns have been the object of extended research, and one of the facts that came out is their frequency. According to Hinds' count of 615 nominal referring expressions in different types of discourse (written/spoken, male/female) only 6 % were overt pronouns whereas 46 % were zero pronouns (Hinds 1983). In Japanese zero pronouns rather than overt pronouns are the primary referring expressions for entities that an utterance most centrally concerns.

Zero pronouns are a typological feature of languages such as Japanese, Korean, Chinese, Thai, Vietnamese, and to point out this typological fact Kameyama has called them 'zero pronominal languages' (Kameyama 1989).

As to the organization of transfer, a transfer component for a pair of zero pronominal languages is supposed to map, in an important number of cases, zero pronouns to zero pronouns. As to Japanese and English, there is a structural difference on the level of pronouns, a typological difference between a 'zero pronominal language' and an 'overt pronominal language'.

The investigations of zero pronouns reported in the literature did not yet concentrate on the role of zero pronouns in a specific task domain, their types, frequency, interpretation procedures. This, however, has been the objective of our case study. But the role of zero pronouns in a VERBMOBIL scenario where a translation system may be used 1 as a supplementary device, has not yet been explored. How pronominal reference is organized in a VERBMOBIL setting with two languages having contrary referring conventions (overt pronouns vs. zero pronouns) we simply do not know.

3 Zero pronouns in the VERBMOBIL domain

We conducted a pilot study on dialogue translation in the VERBMOBIL domain. Dialogues between two participants, one of them Japanese, were translated by a professional Japanese interpreter. We recorded and transcribed 11 dialogues (i.e. about 3 hours of dialogue exchange). The participants were asked to schedule a project meeting, and since the time schedules of both sides were different, the task required search and negotiation. The units translated were turns (i.e. a phrase, a sentence, a short sequence of sentences).

As to zero pronouns, our results may be summarized by several hypotheses, to be confirmed or disconfirmed by further VERBMOBIL related studies.

1. In task-oriented dialogues different forms of domain-specific information seem to be more important for the interpretation of zero pronouns than general context mechanisms on discourse level (e.g. centering (Kameyama 1986; Brennan et al. 1987; Yoshimoto 1988; Walker et al. 1990)).

2. Interpretation procedures for zero pronouns should exploit different sources

of information:

- special domain-dependent 'surface indicators' (types of verbs, morphemes; conventional formulas);
- general dialogue-related information (phases of a dialogue; global goal of the interaction);
- task-related information (e.g. structure and theme of a proposal; memory of proposals made; special problem solving strategies);
- general context mechanisms on discourse level (e.g. centering).

3. As to the ordering of interpretation procedures:

   - local domain-specific indicators should be exploited first;
   - more global information of task-oriented dialogues should take precedence over general context mechanisms on discourse level.

Evidence for these hypotheses will be presented in the following sections.

3.1 Zero pronouns and special domain-dependent indicators

According to our dialogue data, there is a very small group of surface (lexical) indicators relevant for the interpretation of zero pronouns. These indicators occur rather often, especially when time proposals are made, initiated or confirmed (either 'one-sided proposals' or 'common proposals'). These indicators are rather reliable, they are used as reference for speaker or hearer or for both.

Four subgroups may be distinguished (for zero subjects): (cf. next page)

Lexical-Pragmatic-Restrictions

```
(1)   as ano getsuyōbi wa  kyūjitsu nan de    .tabun ano minasan .
             monday    WA  holiday      COP    maybe     everybody

      sanka       dekinai to  omoun desu ga
      participate can not TO  think COP  SAP

      (Monday is a holiday, maybe not everybody can participate,
       (I) think.)


(4)   suijōbi    kara  jikan toritai      desu
      wendsday   from  time  take (want)  COP

      ((I) want to make an appointment on wendsday)

      sō shimashō
      so do PROP
      (let's do it like that)


(2)   nikai     ni   wakemashō ka
      2 parts   NI   di- PROP  QUE
                     vide
      (Shall (we) divide (the time) into two blocks?)

      jikan     chijimete   yarimasen ka (?)
      time      cut         do  NEG   QUE
      (Shall (we) not cut down on time?)


(3)   iie iie asa sa .     sanjikan jikan toremasu .  kuji      kara
      no  no  morning      3 hours  time  can take    9 o'clock from

      jūniji     made no   aida      daijōbu desu
      12 o'clock till NO   between   okay    COP

      (no, no, in the morning, (we) can take three hours off, from 9 to
       12 o'clock would be okay)

(4)   tabun .   kaigi    ni   sanka suru koto   wa  dekimasen
      maybe     meeting  NI   participateKOTO   WA  can   NEG
      (maybe (we) cannot attend the meeting.)
```

(1) a proposal, which speaker-side thinks is possible (or not), is made;
indicators: *(-tai) to omon*
: *mood declarative.*

(2) a proposal, which speaker-side puts forward as a common proposal, is made;
indicators: - *masenka (-mashoka, -masho)*
(stilistic variations according to different settings).

(3) a time proposal is in focus, a standard expression with a special verb is used;
indicators: *toreru*
: *mood declarative* (reference to speaker)
" *interrogative* (reference to hearer)
: *negation* − (proposal made)
" + (proposal not accepted).

(4) Some indicators are less frequent (and less domain-specific). They may be taken as preferences to be further evaluated by other kinds of evidence.

(a) speaker-side wants (doesn't want) something;
indicators: - *tai*
: *mood declarative*
: *negation* − *(+)*

(b) speaker-side is able (not able) to do something;
indicators: *dekiru*
: *mood declarative*
: *negation* − *(+)*

We understand the term 'indicator' as pointing to aspects of a syntactic description. A processing of isolated cues would not be sufficient. Take for example the occurence of a form of dekiru in (1) (list of examples). Since there is already an overt subject, there is no need to apply interpretation procedures for zero subjects. Zero pronouns are found on the basis of syntactic structure, and the indicators are used in a 'default situation': if there is no information to the contrary, i.e. if there are no overt subjects or objects in the utterance, then certain aspects of the syntactic structure function as indicators, and the structure may be completed (if necessary for further processing, transfer included). A formal device for it may be default unification (cf. Bouma 1990; Bouma 1992; Senf & Witt 1994) Another solution would be to use resolution rules which operate on underspecified semantic representations (connected to morphological, syntactic and lexical information) (cf. Alshawy et al. 1992).

Notice that the zero pronouns under discussion refer directly to the discourse situation. There is no anaphoric dependence.

3.2 Zero pronouns and general context procedures

The indicators presented above, though very useful in many cases, are not applicable in some cases, especially when the immediate utterance context has to be taken into account for zero pronoun interpretation:

`e. kayōbi wa watashidomo no tokoro de wa kyūjitsu`

```
     Tuesday  WA   we          NO  side   DE  WA  holiday
  na no de.  anō .  tabun .  ∅ kaigi    ni   sanka suru koto
  COP               maybe      meeting  NI   attend     NOM
  wa  dekimasen
  WA   can   NEG
( On our side Tuesday is a holiday, maybe (we) cannot attend
  the meeting.)
```

This example is interesting in several respects. The subject of *dekimasen* is missing. According to surface indicators (cf. subtype (4)), there is a slight preference for a subject referring to speaker-side. According to principles of centering and their application to zero pronouns, the topicalized element of the preceding utterance is the most prominent antecedent of a zero pronoun. In our case there are two such elements. However, when semantic restrictions are considered, there is a clear preference for a subject referring to speaker-side. Preferences of different sources of information are converging.

There is some evidence in our data that local domain-specific indicators for zero pronouns should be considered first and before general context procedures. The example given on the next page is very interesting in this respect.

The subject of the last verb (a form of toreru) is missing: According to surface indicators (subtype (3)) the subject is expected to refer to speaker-side. According to principles of centering applicable to the immediately preceding utterances, the most prominent antecedent of a zero pronoun would be the subject; in this case 'a collaborator', referred to twice (Kenkyūin, hitori). This example (as well as others found in our data or construed and tested) shows that general context procedures may be not specific enough, and if there are applicable domain-dependent procedures they should be tried first.

```
32
G:              hmm
T:    vormittags
J:              hh ēto watashitachi no tokono kenkyūin no desu ne
                                                  ●
33
G:
T:
J:    e hitori ga desu ne kono hi chotto jūichiji made jikan
           ●                ○
34
G:
T:
J:    fusagatte runde gozenchū to iu koto de aru naraba nijikan

35
G:
T:                                     also leider ein Mitarbeiter
J:    shika jikan ga torenain desu ne

36
G:
T:    von uns äh haben Termin also so einen Termin bis elf Uhr
J:

37
G:
T:    und deshalb vormittags Mittwoch vormittags können wir nur
J:

38
G:
T:    a zwei Stunden zur Verfügung stellen
J:
```

3.3 Zero pronouns and aspects of a task-specific dialogue model

Zero pronouns occurring in task-oriented dialogues may be taken as linguistic evidence or 'diagnostic hints' about task-related context and background knowledge on which speakers rely in naturally occuring dialogues. A formal reconstruction of this dialogue practice has to combine techniques of semantic construction, domain modelling and dialogue processing.

According to our data, different subcases can be distinguished:

(1) Zero pronouns and Idioms

```
A: getsuyō no  gogo       wa   ikaga   desu ka
   monday  NO  afternoon  WA   possibleCOP  QUE
B: ∅ mutsukashii desu.
     difficult   COP
(A: Would monday afternoon be possible?
 B: That would be difficult.)

J: hai wakarimashita
   yes agree
(Yes, (I) see.)
```

Zero pronouns appear in conventional formulas, e.g. of agreement and disagreement.

Another way to express agreement is to repeat the verb of the preceding utterance of the last turn (cf. (4) below). In a dialogue memory this verb-related information should be easily available.

(2) Zero pronouns and general dialogue-related information

```
nihongawa      purojekuto no   mono desu
japanese side  project    NO   MONO COP
                                •
((I) am from the Japanese project group.)
```

Two pieces of information may be assumed to belong to general dialogue-related information:

(a) the global goal of the interaction is to schedule a meeting for members of a Japanese and a German project group;

(b) a dialogue consists of a sequence of phases and phases of a sequence of subtasks.

In our example mono is lexically ambigous referring to objects or persons. Project 'matters' under discussion are persons (according to (a)). The dialogue memory should identify the actual phase as initial phase (context stacks are empty). Self-introduction may be part of the initial phase (according to (b)). One way to introduce oneself is to present oneself as member of a group.

(3) Zero pronouns and thematic structure

```
∅ ∅ jūji       kara  jūsanji    made shika   nain     desu kedo
    10 o'clock from  13 o'clock till besides not have COP

(except from ten a.m.  to one p.m.  (we) have no (time)
```

The object of the verb is missing and not mentioned in the preceding utterance, the subject is missing. Information about the task structure is necessary, specifying: how proposals are made; what they are about; which information is accessible to which dialogue participant.

(4) Zero pronouns and dialogue conventions

```
sumimasen ∅ iu.   ii   wasuremashita ga   sono tōri  desu
sorry       say   say  forget(PAST)  but  the  same  COP

(Sorry, (I) forgot to mention (it), you're right)
```

Subject and object are missing. Information about speech act sequences and speech act realizations is necessary. In the preceding turn speaker A asks: Is it again the case that p (asking for confirmation). In the following turn speaker B presents an excuse and says that he forgot to say that p (giving the confirmation).

A similar example is the following:

```
A: ∅ sanjikan  wa   shikashi  nantoka   torerun   deshō ka
     3 hours   WA   maybe     somehow   can take  COP   QUE
B: ∅ ∅ toremasu
       can take
(A:  Three hours (you) could possibly somehow take off?
 B:  Yes, (I) could.)
```

Speaker A asks for confirmation and speaker B gives the confirmation. According to surface indicators the subject of the first utterance should refer to the hearer.

(5) Zero pronouns and dialogue memory for proposals

```
a .  goji       kara rokuji     made no aida     hitori no hito
     5 o'clock  from 6 o'clock  till NO between  one    NO person

ga korarenain    desu ga .  ano .  su.  suiyōbi   ga jikan
GA come POSS NEG COP  but               wendsday  GA time

toreru    tte itta     no wa sono .  hoka   no hi  no
take POSS TO  say PAST NO WA that    other  NO day NO

gozenchū .  gogo      no aida    sono sanjikan .  jikan ga
morning     afternoon NO in      that 3 hours     time  GA

toreru     to iu imi de ittan desu ne
```

```
take POSS    sense   DE said  COP  NE

(From 5 to 6 o'clock one person cannot come.  When  ? told ? ,
 that  ? would be free on Wendsday, ? meant that ? could take
 three hours off in the morning or in the afternoon)
```

This example is interesting with respect to 'processing insights'. The positions with question marks in the English translation should be filled with expressions referring to the speaker and the hearer, but in which order?

According to surface indicators (subcase 3), the preferred subject candidate should refer to the speaker. But since there is a reference to an earlier proposal this has to be checked by means of a dialogue memory for proposals. If it comes out that the proposal has been made by the hearer, then the subject should refer to the hearer. A default situation does no longer exist and the choice based on surface indicators will be overwritten. This shows that proposals have to be registered (their subject as well as their proposer) and to be accessible for the interpretation of zero pronouns.

(6) Zero pronouns and problem solving strategies

```
demo ∅ mazu   hoka   no.hi    o.   ichiō.mite minnasan
but    first  other  NO day   WO   once  look everybody

jikan ga  atta hō ga ∅
time  GA  have

(But let (us) first have a look to another date, it would be
 (better) if everybody is free)
```

The subject is missing. The utterance is part of a negotiation phase: a proposal for a meeting has been made, but one researcher would be unable to attend. So there is a conflict. One way to proceed is to continue the common search for a better date. Since this is understood, the subject may be zero.

4 Some concluding remarks

Though zero pronouns have been a prominent topic of linguistic or computational linguistic studies, they have hardly been studied with respect to dialogues occurring in a specific task-domain or with respect to translation procedures in such a domain. An analysis of our data leads to the conclusion that general interpretation procedures for zero pronouns in Japanese that have been developed in discourse analysis or text processing are less applicable to dialogues of a task-specific domain (cf. Siegel & Metzing 1994). Therefore, domain-specific interpretation procedures are required. The choice of formalisms for domain-specific interpretation procedures has not been addressed in this paper. The applicability of default-unification to surface indicators has been discussed in (Senf & Witt 1994). The applicability of circumscription to aspects of domain knowledge (task-specific or dialogue structure specific) will be discussed in a subsequent paper. The application of prioritized circumscription (Lifschitz 1989) to interpretation problems of pronouns in English has been explored in recent work by Kameyama, and it seems promising to continue research in this direction.

5 References

Alshawi, H. (ed.) 1992. *The Core Language Engine.* The MIT Press, Cambridge, MA.

Brennan, S., L. Friedman and C. Pollard. 1987. A Centering Approach to Pronouns. In: *Proceedings of the 25th Annual Meeting of the Association for Computatonal Linguistics,* pp. 155–162

Bouma, G. 1990. Defaults in Unifications Grammar, In: *ACL 28*, pp. 165–172.

Bouma, G. 1992. Feature Structures and Nonmonotonicity , In: *Computational Linguistics 18 (2)*, pp. 183–203.

Hinds, J. 1983. Topic continuity in Japanese. In: Givon, Talmy (ed.), *Topic continuity in discourse: a quantitative cross-language study.* Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 43-94.

Kameyama, M. 1985. *Zero anaphora: the case of Japanese.* Stanford University doctoral dissertation.

Kameyama, M. 1986. A Property-sharing Constraint in Centering. In: *Proceedings of the 24th Annual Meeting of the Association of Computational Linguistics.* New York, NY, pp. 200-206.

Kameyama, M. 1989. *Functional Precedence Conditions on Overt and Zero Pronominals.* Ms., MCC April 1989.

Kameyama, M. 1994. *Indefeasible Semantics and Defeasible Pragmatics.* AI Center and CSLI, Menlo Park.

Lifschitz, V. 1989. Circumscriptive Theories. In: R. Thomason (ed.), *Philosophical Logic and Artificial Intelligence.* Reidel, pp. 109-159.

Senf, Th. and A. Witt. 1994. *Der Nutzen von HPSG-Satzrepräsentationen für die Bestimmung von Antezedenten der Nullpronomina.* VERBMOBIL-Memo 20, University of Bielefeld.

Siegel, M. 1993. *Dialogdolmetschen. Eine Pilotstudie zu aufgabenorientierten Dialogen (Terminabsprachen) Japanisch-Deutsch.* Arbeitsberichte Computerlinguistik 3-93, University of Bielefeld.

Siegel, M., H. Kuroda and E. Kubo. 1993. *Dialogdaten: Terminplanung Japanisch-Deutsch.* Arbeitsberichte Computerlinguistik 2-93, University of Bielefeld.

Siegel, M. and D. Metzing. 1994. *Nullpronomina und die Organisation von Wissensquellen für den Transfer Japanisch - Englisch.* VERBMOBIL-Memo 12, University of Bielefeld.

Walker, M., M. Iida and Sh. Cote. 1990. Centering in Japanese Discourse. In: *Proceedings of Coling '90.*

Yoshimoto, K. 1988. Identifying Zero Pronouns in Japanese Dialogue. In: *Proceedings of Coling 88.* Budapest, pp. 779-784.