"TRANSLATION GREAT PROBLEM" - ON THE PROBLEM OF INSERTING ARTICLES WHEN TRANSLATING FROM RUSSIAN INTO SWEDISH

By Barbara Gawrońska-Werngren
Dept of Linguistics, Lund University, SWEDEN
Helgonabacken 12, S-22362 LUND, e-mail: linglund@gemini.ldc.se

The problem to be discussed here - i.e. how to generate exponents of a morphosyntactic feature which is systematically used in the target language, but not in the source language - is closely related to the development of SWETRA - a multilanguage MT system for translating between fragments of Russian, Swedish, English and German (Sigurd & Gawrońska-Werngren, 1988). Anyone working on translation between Russian and Germanic languages must face difficulties as Russian NPs do not have either indefinite or definite articles.

The solutions proposed here have been implemented in the SWETRA - program, which is based on a functional GPSG formalism called Referent Grammar (RG; Sigurd 1987). RG-rewriting rules, implemented in Definite Clause Grammar, are used both for analysis and synthesis. The result of parsing is a so-called functional representation (f-representation), containing descriptions of the constituents and information about their syntactic functions. An f-representation of a simple transitive sentence like "a boy met a girl" looks like this:

```
s(subj(np(r(_,m(boy,sg),indef,sg,_,_),
      Attr1,Relcl1))
pred(m(meet,past)),
obj(np(r(_,m(girl,sg),indef,sg,_,_,_)
         Attr2,Relcl2)),
sadvl([]),sadvl([]),advl([]),
advl([]),advl([])).
```

The entity with the functor r, called "referent nucleus", is a description of the head noun. Slots Attr1/Attr2 and Relcl1Relcl2 are used, respectively, for storing possible attributes and relative clauses.

Given an instantiated f-representation, the program can generate the target equivalent of the input string according to target-specific rules. But if a certain value required in the target language (as definiteness in Swedish and English) is unspecified in the source language (as definiteness in Russian), the information stored in the f-representation may be insufficient for generating a grammatically correct output (although the output may be comprehensible). So there is a need of an intermediate (transfer) stage between analysis and synthesis. The most probable definiteness values must be derived from the context before the target rules for marking definiteness start to work. Since the notions of reference and co-reference are crucial when choosing definiteness values, this intermediate stage will be called "referent tracking".

A preliminary discourse model for referent tracking

Informally, discourse referents are often defined as "things the sender is talking about". Referring means primarily pointing out objects and facts in the external world, but we have also to pay attention to those linguistic factors which enable identifying two or more phrases as co-referential. Obviously, two co-referential words or strings of words do not have to point out a physically existing thing: they may allude to an event or an abstract concept. So discourse referents must be understood as cognitive entities existing in the mental world.

In the program for referent tracking discussed below, a distinction is drawn between nominal referents - alluding to objects: cats, unicorns etc. - roughly, to things which can be pointed out by non-linguistic means, in potential (unicorns may be pointed out on a picture or drawn) and "event referents" - referents of whole predications or predicative (verbal) NPs. "Event referents" correspond to situations, actions or relations between objects. This distinction is not unproblematic (there are obviously borderline cases), but it is useful for translation purposes, since definiteness may be triggered not only by an NP, but also by a predication as a whole. As will be shown below, the rules for discovering co-reference have to be formulated in different ways depending on which kind of referent (nominal referents or events) is involved.

Referent tracking and generation of definiteness values

A model for generating definiteness cannot be based on the simplistic principle: if an NP with a given meaning has been translated previously (in the current text), provide it with the value "definite"; otherwise, treat it as indefinite. In order to instantiate the definiteness value, we have to investigate the internal structure of the NP, the interplay between the current NP and the other syntactic constituents of the analyzed sentence as well as the relations between the current NP and the previously translated part of the text.

The preliminary procedure inserting definiteness values used in the RG-model contains the following stages:

A. Investigating the functional representation of the first sentence of the input text in order to create a "preliminary discourse frame".

B. Storing the descriptions of noun phrases (including their referent numbers) and representations of "events" in a data base.

C. Comparing the representations of noun phrases in the current sentence with the stored information in order to discover possible co-reference; storing new "events" and new "nominal referents", if any.

The right noun phrase form is then generated according to language specific rules - e.g. rules which do not allow NPs like **the my book* or Swedish **min boken* (my book+def) and rules inserting possessive pronouns before nouns denoting close relationship, like "brother", "neighbour" etc. A Russian sentence like*Ja vstretil soseda* (I met neighbour) is translated into Swedish as

Jag träffade min granne
I met my neighbour.

Stage A includes subprocedures like:

- checking if the current sentence is a predicative construction as "X is a great linguist"; if yes - the referent representation of X has to be provided with the attribute meaning "great linguist" before storing in order to enable co-reference identification in the later part of the text, where X may be referred to by an NP like "this great linguist".
- checking whether the sentence contains specific time and/or place adverbials, whether the current NP contains any attributes which may be interpreted as definiteness indices and whether there are any constituents having clearly specific reference. The aim is to classify the current NP and the whole predication as to their reference: if the sentence evokes many

specific concepts and/or the NP contains reference restricting attributes, we may assume, that the event referred to is highly specific, and that the probability for definite articles may become greater (if no counterindices can be found). The results are not always plausible and can probably be improved by more work on topic - comment relations. Currently, when translating a sentence fragment like:

včera večerom Michail Gorbačev
yesterday evening Michail Gorbachev
vydvinul predloženie ob...
made proposal about

the program inserts the value "prodef" (probably definite) in the representation of the noun meaning "proposal", as the discourse frame is highly specific: it contains a specific time value, a specific subject referent and a specification of the noun meaning "proposal" by means of a prepositional phrase. Thus, the Swedish translation version below gets greater preference:

igår kväll lade Michail Gorbatjov
yesterday evening put Michail Gorbachev
fram förslaget om...
forward proposal+def about

although many native speakers of Swedish would prefer the alternative variant:

igår kväll lade Michail Gorbatjev
yesterday evening put Michail Gorbachev
fram ett förslag om...
forward a proposal about

The second variant is of course not excluded by the subprocedure. Nevertheless, even if the first output is not always the most preferred one, checking the degree of specificity is often useful. If we deleted this part of the translation procedure, every NP in the first sentence of a text would be understood as indefinite, something which would lead to many "strange" translations (*a professor at a department of linguistics at a university of Lund*).

If the first sentence in the text does not contain any definiteness indices, the definiteness slot remains anonymous and gets the default value "indef(inite)" during the generation process, if no target-specific rules prevent it.

The information supported by the sentence is stored in two lists: a "nominal referent list" - for characteristics of those NPs which have been interpreted as establishing nominal referents, and an "event list", where representations of predications (including those expressed by verbal nouns) are placed. Each new NP to be translated is now compared with the stored information - the aim is to discover possible definiteness triggers. The simplest case of definiteness triggering is that of nominal co-reference (the current NP points out a nominal referent which has been introduced before). Nevertheless, a procedure handling this "simple" case must be quite elaborated, as it has to cover at least the following cases:

- co-reference between NPs with identical head nouns: here, the program must check if the current NP contains attributes which exclude co-reference with a previously translated NP having the same head-meaning code. In a sequence like *A boy played with a little dog. Then, a big dog came* the two dogs must not be interpreted as co-referential. This is achieved by a subprocedure "attribute_conflict", which compares the attributes of the NPs involved.
- co-reference between synonyms or between a hyponym and a hyperonym: the program must be able to trigger the value "prodef" if the current NP evokes a concept which is not more

restricted than and not incompatible with a previously stored referent. Thus, the strings *my old teacher* and *man* should be identified as co-referential in a sequence like: *I met my old teacher. The man was drunk*; but not in *I met a man. My old teacher was drunk.*. Furthermore, if the current NP refers to a set of objects, we have to check if there are at least two previously established referents which - treated as elements of a set - constitute a potentially co-referential set (cases like: *A boy met a girl. The children ran home*). For this purpose, recursive PROLOG-predicates searching for possible hyponyms in the referent list are used. One of the simpler versions of the predicate for co-reference discovery (the one handling cases like boy+girl=children+def) is formulated as follows:

```
possible_coref(m(A,pl),Rlist):-
 hyponyms(m(A,sg),[H|T],Rlist).
```

where m(A,pl) is the meaning code of the current noun, Rlist is a list containing codes of previously translated noun phrases and the possible hyponyms of the singular form meaning A are stored in the list [H|T]. The whole rule is to be read as: a plural noun with the meaning code m(A,pl) may co-refer with a set containing referents of previously mentioned NPs, if at least two previously mentioned nouns can be interpreted as hyponyms of the singular form of the current noun. The predicate "hyponyms" utilizes the semantic features stored in lexical entries in order to establish a hierarchy between meaning codes.

- co-reference between evaluating and non-evaluating expressions - as in the following fragment of a Pravda-notice:

Israeli airplanes staged three bomb-attacks on Lebanese territory today.

Fifteen persons were killed as a result of the barbaric action of the air pirates.

The evaluation of israeli airplanes as "air-pirates" depends obviously on the sender's attitude, and such aspects as the sender's political and emotional preferences are not accessible to the program. But evaluating components seem not to restrict the potential reference of an NP in a purely linguistic way (any human being may be referred to by an NP like *this fool*). Therefore, we may assume, that if the general condition for possible co-reference (not incompatible and not more restricted) is fulfilled after extraction of evaluating elements from the semantic characteristics of the current NP, definiteness may be triggered. In the example above, after deleting evaluations from the lexical description of the entity "air-pirate", the features corresponding to the concepts "airplane" and "pilot" remain. Consequently, co-reference with "israeli airplanes" is not excluded.

- whole - part relations: in cases like car - engine etc. definiteness should be triggered. Formulating a PROLOG-rule handling this kind of relation is not a difficult task - the problem is to create an appropriate data base (it would be necessary to include much encyclopaedic knowledge in the lexicon).

Another type of definiteness triggering rules applies in the case of co-reference between sequences alluding to "events", as in the following example:

An unidentified submarine followed a Swedish trawler.

The hunt went on for about two hours.

The first step is to check whether the current noun (here: *hunt*) may be interpreted as having an "event-referent" - the information is provided in the lexicon. Then, a specific rule for possible event-co-reference applies. It would not be sufficient to compare the semantic representation of "hunt" with that of the finite verb ("follow") according to the previously outlined principle: "not incompatible and not more restricted". "Hunting" is obviously a more specific concept than following (hunting is a special type of following). As the NP meaning "hunt" refers to an event, we have to treat it as a predication and compare it with the previously mentioned predication as a whole. The event-list contains at this point a representation formulated as:

```
e(hunt,args(r(1,submarine,unidentified),
            r(2,trawler,swedish)))
```

The event referred to by *hunt* has no syntactically represented arguments - before co-reference checking it gets a representation like: e(hunt,args(_,_)). Co-reference seems to be allowed by the following principle: a verbal noun may co-refer with a previous predication, if it is semantically not incompatible with the predicate and if the arguments of the verbal noun are either not specified or co-referential with the arguments of the previously stored predicate. A PROLOG-implementation of this rule may have the following shape (simplified):

```
possible_coref(NewEvent,OldEvent):-
NewEvent= e(m(Mean,verbal),args(A1,A2)),
OldEvent = e(Pred,args(A3,A4)),
eventlist(Elist), member(OldEvent,Elist),
not(incompatible(Mean,Pred)),
(var(A1);possible_coref(A1,A3)),
(var(A4);possible_coref(A2,A4)).
```

The case of "pseudo-objects"

In the example above, both syntactic arguments of the transitive verb were clearly referential - they pointed out specific objects. But there are cases in which the syntactic complement of a verb does not allude to a referent - though the form of the complement is nominal. The distinction is manifested clearly in Swedish, where the stress pattern of the string verb + complement varies depending on whether the complement is referential or non-referential. In the second case, the stress pattern is identical with the one of particle verbs. Furthermore, the complement cannot take relative clauses:

i. han höll *tal*
he made speech

ii.**han höll tal som var fint*
he made speech that was fine

If *höll* takes an object proper, as in iii., the stress pattern changes:

iii.*han höll ett (långt) tal som var fint*
he made a (long) speech that was fine

The unability versus ability of taking relative clauses is highly significant and can be taken as a criterion for referent establishment. According to RG (Sigurd 1989), the head noun, the relative pronoun and the relativized (lacking) constituent in the subordinate (defective) clause are considered as alluding to the same referent. The ungrammaticality of relative clauses other than sentence relativizing ones can be explained by the fact that the "pseudo-object" *tal* lacks a referent of its own. The only accessible referent which can be common for the relative pronoun and the lacking constituent in the relative clause is the referent of the whole predication - as in iv.:

iv. *han höll tal vilket var fint*
he made speech which was fine

Vilket is the only Swedish pronoun used for sentence relativization. The sentence above may be paraphrased as: *det var fint att han höll tal* ('it was fine that he made a speech') or as *att* *han höll tal var fint* ('that he made a speech was fine') but not as **han höll tal som var fint* ('he made a speech that was fine'). Subsequently, components which cannot contain relative clauses are treated as incapable of establishing referents of their own. In the referent tracking procedure, they are interpreted as components of the verbal part of an event. The translation problem arising here is caused by the fact that the distinction between referential objects and "pseudo-objects" is not manifested in Russian. Both v. and vi. are possible:

v. *on proiznes reč'*
he "made" speech

vi. *on proiznes (dlinnuju) reč', kotoraja*
he made (long) speech that
nikomu ne ponravilas'
nobody+dat not liked

v. may thus be translated into Swedish either as *han höll tal* or *han höll ett tal*.. This translation procedure preserves the ambiquity. If there are neither relative clauses nor other attributes before/after a form which may be interpreted as a "pseudo-object", and if there are no counterindices (e.g.clearly anaphoric expressions in the next following part of the text) the non-referential interpretation is preferred, but the second alternative (*han höll ett tal*) is not excluded.

Summary

The model and procedures discussed above are attempts to utilize text semantic restrictions in machine translation. The current version of the program covers quite a large repertoire of different types of definiteness-triggers and handles generation of correct forms of "pseudo-objects" in phrases like "play the piano", "play footboll" etc. quite successfully. Nevertheless, there is a need for further study - among other problems, on the "life-span" of discourse referents and on cases where NPs traditionally (i.e. according to Karttunen 1976) treated as non-referential (e.g. predicatives) allow certain instances of definite anaphora (Frarud 1986). The semantic representations of lexical entries require elaboration, and storing non-linguistic knowledge necessary for appropriate definiteness triggering is a problem. Currently, the program works quite efficiently when translating short text fragments, where the number of discourse referents is not too great.

References:

Frarud, K. 1986. The introduction and maintenance of discourse referents. In: Papers from the 9th Scandinavian Conference of Linguistics, 11-122.

Karttunen, L. 1976. Discourse referents. In: Syntax and Semantics, vol. 7, 383-386. New York: Academic Press.

Sidner, C. L. 1983. Focusing in the comprehension of definite anaphora. In: Brady, M & M. C. Berwick: Computational models of discourse, 267-330, Massachusetts.

Sigurd, B. 1987. Referent Grammar. A Generalized Phrase Structure Grammar with built-in referents.Studia Linguistica 41:2, 115-135.

Sigurd, B. & B. Gawrońska-Werngren. 1988. The Potential of Swetra - A Multilanguage MT System. Computers and Translation 3, 237-250.

Sigurd, B. 1989. A referent grammatical analysis of relative clauses. Acta Linguistica Hafniensia 21:2, 95-115.